Integration of beliefs and affective values in human decision-making

Marion Rouault

Ecole Normale Supérieure

Ecole doctorale Cerveau, Cognition, Comportement

Thèse de Doctorat

pour obtenir le titre de

Docteur en Science

de l'Ecole Normale Supérieure de Paris

Spécialité : Neurosciences

Présentée par

Marion Rouault

Integration of beliefs and affective values in human decision-making

Laboratoire de Neurosciences Cognitives,

Soutenue le 22 septembre 2015

Devant le jury composé de :

Directeur de thèse	Etienne Koechlin	DR INSERM
Rapporteur	Timothy Behrens	PR, University of Oxford
Rapporteur	Emmanuel Procyk	DR CNRS
Examinateur	Boris Burle	DR CNRS
Examinateur	Christian Lorenzi	PR ENS
Examinateur	Mathias Pessiglione	DR INSERM

I, Marion ROUAULT, declare that this thesis titled, 'Integration of beliefs and affective values in human decision-making' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this university.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. with the exception of such quotations, this thesis is entirely my own work.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

signed:

date:

"Il est certains esprits dont les sombres pensées
Sont d'un nuage épais toujours embarrassées ;
Le jour de la raison ne le saurait percer.
Avant donc que d'écrire, apprenez à penser.
Selon que notre idée est plus ou moins obscure,
L'expression la suit, ou moins nette, ou plus pure.
Ce que l'on conçoit bien s'énonce clairement,
Et les mots pour le dire arrivent aisément."

Nicolas BOILEAU (1636-1711)

ECOLE NORMALE SUPERIEURE DE PARIS

Abstract

Laboratoire de Neurosciences Cognitives
INSERM U960

Philosophiæ Doctor. Speciality: Neurosciences

Integration of beliefs and affective values in human decision-making

by Marion Rouault

Executive control relates to the human ability to monitor and flexibly adapt behavior in relation to internal mental states. Specifically, executive control relies on evaluating action outcomes for adjusting subsequent action.

Actions can be reinforced or devaluated given affective value of outcomes, notably in basal ganglia and medial prefrontal cortex. Additionally, outcomes convey information to adapt behavior in relation to internal beliefs, involving prefrontal cortex. Accordingly, action outcomes convey two major types of value signals: (1) Affective values, representing the valuation of action outcomes given subjective preferences and stemming from reinforcement learning; (2) Belief values about how actions map onto outcome contingencies and relating to Bayesian inference. However, how these two signals contribute to decision remains unclear, and previous experimental paradigms confounded them. In this PhD thesis, we investigated whether their dissociation is behaviorally and neurally relevant.

We present several behavioral experiments dissociating these two signals, in the form of probabilistic reversal-learning tasks involving stochastic and changing reward structures. We built a model establishing the functional and computational foundations of such dissociation. It combined two parallel systems: reinforcement learning, modulating affective values, and Bayesian inference, monitoring beliefs. The model accounted for behavior better than many other alternative models.

We then investigated whether beliefs and affective values have distinct neural bases using fMRI. BOLD signal was regressed against choice-dependent and choice-independent beliefs and affective values. Ventromedial prefrontal cortex (VMPFC) and midcingulate cortex (MCC) activity correlated with both choice-dependent variables. However, we found a double-dissociation regarding choice-independent variables, with VMPFC encoding choice-independent beliefs, whereas MCC encoded choice-independent affective values. Additionally, activity in lateral prefrontal cortex (LPFC) increased when decision values (i.e. mixture of beliefs and affective values) got closer to each other and action selection became more difficult.

These results suggest that before decision, VMPFC and MCC separately encode beliefs and affective values respectively. LPFC combines both signals to decide, then feeds back choice information to these medial regions, presumably for updating these value signals according to action outcomes. These results provide new insight into the neural mechanisms of decision-making in prefrontal cortex.

ECOLE NORMALE SUPERIEURE DE PARIS

Abstract

Laboratoire de Neurosciences Cognitives
INSERM U960

Philosophiæ Doctor. Speciality: Neurosciences

Integration of beliefs and affective values in human decision-making

by Marion Rouault

Le contrôle exécutif de l'action fait référence à la capacité de l'Homme à contrôler et adapter son comportement de manière flexible, en lien avec ses états mentaux internes. Il repose sur l'évaluation des conséquences des actions pour ajuster les choix futurs.

Les actions peuvent être renforcées ou dévaluées en fonction de la valeur affective des conséquences, impliquant notamment les ganglions de la base et le cortex préfrontal médian. En outre, les conséquences des actions portent une information, qui permet d'ajuster le comportement en relation avec des croyances internes, impliquant le cortex préfrontal. Ainsi, les conséquences des actions portent deux types de signaux : (1) Une valeur affective, qui représente l'évaluation de la conséquence de l'action selon les préférences subjectives, issue de l'apprentissage par renforcement ; (2) Une valeur de croyance, mesurant comment les actions correspondent aux contingences externes, en lien avec l'inférence bayésienne. Cependant, la contribution de ces deux signaux à la prise de décision reste méconnue. Dans cette thèse, nous avons étudié la pertinence de cette dissociation aux niveaux comportemental et cérébral.

Nous présentons plusieurs expériences comportementales permettant de dissocier ces deux signaux de valeur, sous la forme de tâches d'apprentissage probabiliste avec des structures de récompense stochastiques et changeantes. Nous avons construit un modèle établissant les fondations fonctionnelles et computationnelles de la dissociation. Il combine deux systèmes en parallèle : un système d'apprentissage par renforcement modulant les valeurs affectives, et un système d'inférence bayésienne modulant les croyances. Le modèle explique mieux le comportement que de nombreux modèles alternatifs.

Nous avons ensuite étudié, en IRM fonctionnelle, si les représentations dépendantes et indépendantes du choix des croyances et des valeurs affectives avaient des bases neurales distinctes. L'activité du cortex préfrontal ventromédian (VMPFC) et du cortex mid-cingulaire (MCC) corrèle avec les deux variables dépendantes du choix. Cependant, une double-dissociation a été identifiée concernant les représentations indépendantes du choix, le VMPFC étant spécifique des croyances alors que le MCC est spécifique des valeurs affectives. En outre, l'activité du cortex préfrontal latéral augmente lorsque les deux valeurs de décision sont proches et que le choix devient difficile.

Ces résultats suggèrent qu'avant la décision, le cortex préfrontal ventromédian (VMPFC) et le cortex mid-cingulaire (MCC) encodent séparément les croyances et les valeurs affectives respectivement. Le cortex préfrontal latéral (LPFC) combine les deux signaux pour prendre une décision, puis renvoie l'information du choix aux régions médianes, probablement pour actualiser les deux signaux de valeur en fonction des conséquences du choix. Ces résultats contribuent à élucider les mécanismes cérébraux de la prise de décision dans le cortex préfrontal.

Remerciements

Ces années de thèse sont loin d'avoir été un travail de recherche solitaire et isolé, aussi suis-je ravie d'avoir de nombreux remerciements à exprimer !

Je souhaite remercier chaleureusement Etienne Koechlin d'avoir accepté d'encadrer ce travail de thèse. Je lui suis très reconnaissante de la confiance et de l'indépendance qu'il m'a accordées, ainsi que de sa patience. Sa passion pour les neurosciences aura été contagieuse ! Je lui exprime également ma gratitude pour tout ce qu'il m'a transmis de sa conception du travail de recherche. J'espère avoir hérité de sa rigueur scientifique et de son exigence intellectuelle. J'ai apprécié nos nombreuses discussions scientifiques et extra-scientifiques, qui j'espère se poursuivront bien au-delà de cette thèse.

Je remercie les membres de mon jury d'avoir accepté d'évaluer ce travail, en particulier car j'imagine qu'il existe lectures plus agréables en plein été : Timothy Behrens, Boris Burle, Christian Lorenzi, Mathias Pessiglione et Emmanuel Procyk.

Je remercie l'Ecole Normale Supérieure de Lyon de m'avoir permis de faire de longues études et de m'avoir financée pendant mon doctorat. Je remercie les nombreux volontaires qui ont accepté de participer à mes expériences, ainsi que les membres du Centre de Neuroimagerie de Recherche de la Pitié-Salpêtrière et Alice, qui ont passé de longues heures à m'aider à enregistrer les données d'IRM.

Je souhaite remercier toutes les personnes de l'équipe "Frontal Lobe Functions", qui, tour à tour, ont alimenté une atmosphère amicale. Grâce à eux, ces quatre années auront été très stimulantes sur les plans scientifique et intellectuel : Charles, Gabriel, Héloïse, Jan, Maël, Muriel, Philippe, Stefano, Sven et Valentin. Je remercie également Laura et Marine pour leur sympathie et leur efficacité. Un merci très chaleureux à Anne-Do et à Valérian pour leur bienveillance et leurs nombreux conseils. Enfin, je remercie les autres personnes qui ont fait du labo un environnement joyeux et amical : Emma, Flora, Florence, Guillaume, Mariana, Marwa et Thibaud.

Je remercie particulièrement Guillaume de m'avoir fait, le premier, découvrir les neurosciences cognitives, et de m'avoir encouragée à explorer ma propre voie par la suite. Je remercie profondément Marie-Hélène de son optimisme et de sa gentillesse.

Je termine en exprimant ma gratitude aux personnes qui, de près ou de loin, m'ont entourée pendant ces années de thèse, et dont l'amitié m'est précieuse : Amélie, Blandine, François et ses idées brillantes, Julie et Florence, pour leur justesse et leur perspicacité, Laura, pour son entrain et son optimisme, Maïlys, ainsi que les colocs qui ont partagé avec moi ces années parisiennes : Fabian, Quentin et Aurélien, sans oublier Adrien et Lucie.

J'ai la chance d'être entourée d'une famille géniale. Je tiens à remercier infiniment mes parents pour leur amour inconditionnel et leur soutien constant. Ils sous-estiment le fait que c'est aussi grâce à eux que j'ai pu arriver aujourd'hui là où je souhaitais arriver. Je remercie aussi mes deux petites soeurs : Pauline et Alice, de leur présence et de leurs petites attentions. Promis, quelque soit la distance géographique, nous resterons proches !

Enfin, je remercie Clément, pour tout !

Paris, le 16 juillet 2015

Contents

Abbreviations

ACC	**A**nterior **C**ingulate **C**ortex
AIC	**A**kaike **I**information **C**riterion
BIC	**B**ayesian **I**information **C**riterion
BA	**B**rodmann **A**rea
dACC	**d**orsal **A**nterior **C**ingulate **C**ortex
dmPFC	**d**orso **m**edial **P**re **F**rontal **C**ortex
ERN	**E**rror **R**elated **N**egativity
fMRI	**f**unctional **M**agnetic **R**esonance **I**maging
FPC	**F**ronto **P**olar **C**ortex
GLM	**G**eneral **L**inear **M**odel
LPFC	**L**ateral **P**re **F**rontal **C**ortex
LFP	**L**ocal **F**ield **P**otential
LLH	**L**og-**L**ikely**H**ood
MCC	**M**id **C**ingulate **C**ortex
MNI	**M**ontreal **N**eurological **I**nstitute
OCD	**O**bsessive **C**ompulsive **D**isorder
OFC	**O**rbito **F**rontal **C**ortex
PCC	**P**osterior **C**ingulate **C**ortex
PFC	**P**re **F**rontal **C**ortex
RL	**R**einforcement **L**earning
ROI	**R**egion **O**f **I**nterest
SMA	**S**upplementary **M**otor **A**rea
tDCS	**t**ranscranial **D**irect **C**urrent **S**timulation
TMS	**T**ranscranial **M**agnetic **S**timulation

TPJ	**T**enporo **P**arietal **J**unction
vmPFC	**v**entro **m**edial **P**re **F**rontal **C**ortex

Foreword

Decision-making is a critical feature for survival, in a permanently evolving environment. In humans, decisions are considered to be the ultimate expression of free will and voluntary behavior. Human behavior is characterized by an important flexibility and adaptability, two elements which allow humans to realize their internal goals through the decisions they make. This flexible adaptability is crucial especially given that everyday decisions take place in ever-changing environments. Accurately evaluating the value of choice options at stake is therefore critical. In this PhD work, we investigated the outcome evaluation mechanisms underlying free choice in sequential decisions, two features that are close to real-life choices.

For centuries, philosophers, psychologists and economists have tempted to access our internal world through the means of introspection and the study of behavior. In the past decades, functional magnetic resonance imaging has revolutionized the study of human brain mechanisms. Despite providing only correlational evidence, it is a non invasive method allowing to investigate the neural bases of certain cognitive processes or variables. One of the key feature of this PhD work lies in the complementary contributions of experimental and computational approaches to the study of choice cerebral mechanisms.

Decisions manifest themselves through actions but can be dissociated from them. Decisions owe their existence to mental processes hidden within the brain foldings. It seems now established that human medial prefrontal cortex is a key hub in the decision-making network. However, a structural and functional refinement remains to be elaborated. Combining modern approaches such as behavioral psychophysics, computational modeling and neuroimaging, it is now possible to investigate the neural mechanisms underlying decision-making, in order to determine the hidden variables that link perceived outcomes to actions. These hidden variables, which govern subjects' decisions, constitute the interface between the real world and its mental representation.

Chapter 1

Human Decision-Making and Prefrontal Function

My PhD work falls within the general framework of human prefrontal executive function, with a focus on value-based decision-making.

1.1 Prefrontal cortex subserves central executive function

Executive control relates to the human ability to monitor and adapt behavior in relation to internal mental states (Miller and Cohen, 2001 [1]). Indeed, human subjects are able to not respond only to immediate stimuli, in an automatic manner, but also to respond to stimuli in relation to internal goals and beliefs, in an *adaptive* and *flexible* manner. Thus, executive control refers to a set of functions giving humans their aptitude to react not only automatically to external events, but also regarding inner thoughts and intentions, which manifest themselves through desires, objectives and beliefs. These functions are qualified as *central* as they are involved in our perception of ourselves as autonomous and responsible agents, with voluntary intentions. Finally, executive functions are associated with the notion of consciousness. In humans, prefrontal cortex subserves central executive function (Figure 1.1).

1.1.1 Early insights into prefrontal cortex functions: the contribution of lesion studies

Originally, executive control was considered as a set of abilities such as planning, organization and goal-directed behavior, that are implicated a lot in daily life: decision-making,

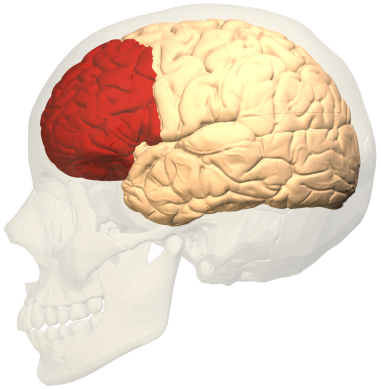


FIGURE 1.1: Prefrontal cortex subserves human central executive function (lateral view, wikipedia image).

reasoning, social interactions, etc. Historically, the first insights into prefrontal function came from a memorable patient named Phineas Gage. Following a serious accident in 1848, Gage survived but was injured in both frontal lobes. Consequently, he sustained temporary behavioral and personality changes (e.g. more impulsivity), along with social skills modifications, directly relating for the first time personality traits with a specific brain region change. The case was later re-studied using modern techniques (Damasio et al., 1994 [2]). Damasio and colleagues confirmed that the deep changes in Phineas Gage's personality were related to damage in both left and right anterior parts of PFC, causing deficit in rational decision-making and emotion processing.

More broadly, patients with prefrontal cortex lesions appear to have general motor, sensory and memory functions preserved, but are seriously impaired in real-life functioning (Shallice and Burgess, 1991 [3]). In two real-life settings tasks, Shallice and Burgess reported 3 patients cases for which they showed deficits in prefrontal function implying an inability to switch between tasks [3]. This result was replicated by other research groups (Rubinstein et al., 1994 [4]). Patients had difficulty interrupting ongoing behavior to execute a different course of action, as well as going back to the original course of action afterwards. This dysexecutive syndrome can present with two main clinical pictures, with variations according to the spatial extent of lesions and to the patient's life history. (1) The hypoactive form is characterized by a lack of initiative, apathy, inertia, and difficulty making decisions. (2) The hyperactive form is characterized by impulsivity, inappropriate behavior, lack of insight on one's own behavioral outcomes, and frequent change of goals. This supports the view that prefrontal cortex subserves auto-regulation and action control abilities.

A more recent review by Szczepanski and Knight (2014) [5] provides a finer characterization of prefrontal lesions regarding the functional specificities of each subregion. In particular, Azuar and colleagues demonstrated that LPFC regions' integrity was necessary to exert cognitive control. Furthermore, the posterior regions integrity was necessary for the most anterior regions to exert such control (Azuar et al., 2014 [6]). These lesion studies are particularly interesting since they provide *causal* relationships between brain area and function, whereas fMRI provides correlative data.

1.1.2 Prefrontal cortex neuroanatomy, cytoarchitecture and neurophysiology

The frontal lobes are particularly developed in humans compared to other species. They form a third of the brain surface and correspond to its most anterior part, incorporating both hemispheres. In this section, we present prefrontal cortex subdivisions according to anatomical landmarks. Functional subdivisions will be discussed in the next section.

Prefrontal cortex is delineated caudally by motor cortex. Premotor cortex and supplementary motor area (SMA) are usually not considered part of prefrontal cortex. Gyri and sulci, giving the human brain its characteristic folded appearance, constitute anatomical landmarks to decompose prefrontal cortex into distinct subparts. However, a decomposition based on Brodmann areas, which is not inconsistent with gyri and sulci, is more often used.

Brodmann areas (BA) give subregions delineation given cytoarchitecture, which refers to the cellular properties of the neural networks composing the different cortical layers. This classification is therefore based on the apparent structural organization of the cortex: number and thickness of cortical layers, dendritic arborization, etc. Figure 1.2 display the Brodmann areas composing prefrontal cortex.

The lateral part of prefrontal cortex comprehends, from rostral to caudal: BA 47 (OFC, ventrally); BA 10 (frontal pole, the most anterior part); BA 46 and BA 9 (roughly corresponding to dorsolateral PFC); BA 8, including frontal eye field (Figure 1.2). On the left hemisphere, BA 44 and BA 45 (inferior frontal gyrus) correspond to Broca area, an area necessary for speech production, that is triggered during semantic tasks, semantic working memory and retrieval, as well as phonological and syntactic processing. The medial part comprises BA 24 (ventral anterior cingulate), BA 25 (subgenus, governing amygdala, insula and hippocampus), BA 32 (dorsal anterior cingulate) and BA 33 (pregenual cingulate). Here, I would like to emphasize the importance of these anatomical landmarks to study functionality. Indeed, usually computational models and

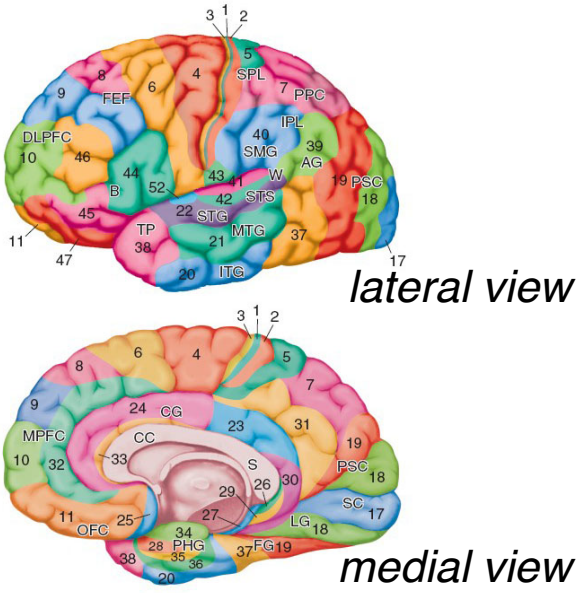


FIGURE 1.2: Prefrontal cortex includes Brodmann areas 8, 9, 10, 11, 12, 24, 25, 32, 33, 44 and 45 (Broca), 46 and 47. These delineations are based on cytoarchitecture (image from Traite de neuropsychologie clinique by Lechevalier and colleagues, 2008).

neuroimaging do not take into account anatomical bases to elucidate brain subregions' function.

Neurotransmitters and connectivity. Cortical layers are composed of excitatory and inhibitory neurons, that have long distance reciprocal projections with the rest of the cortex. Prefrontal cortex presents high intrinsic connectivity, as well as extrinsic connectivity with other brain regions. All neuromodulators types are present in prefrontal cortex (Fuster, 1988 [7]). Specifically, dopamine and norepinephrine, thought to mediate learning (Collins and Frank, 2012 [8]), are found in higher concentrations than in other brain regions. Prefrontal cortex also has glutamatergic projections to the limbic system, e.g. amygdala and hippocampus, which are modulating emotional and memory-related responses, as well as neurons projecting to the thalamus and hypothalamus. Mutual connections i.e. that feed the PFC and that the PFC feeds involve sensory areas and posterior associative areas, making prefrontal cortex a center of convergence for various sensory inputs.

The connections pattern was originally investigated using tracers injection in non-human primates (Petrides and Pandya, 2002 [9]). Today, diffusion tensor imaging (DTI) allowed to uncover part of these tracks (Croxson et al., 2005 [10]) and dress parallels with non-human primates functional regionalization.

1.1.3 Prefrontal cortex in non-human primates and other species

We will see in the next chapters that a lot of what we know about brain structure and function derive from the contribution of animal studies. Although my PhD work concerns the human brain, this section is a complement concerning animal brain anatomy.

Non-human primates brain share homologies with human brain regions (Wise et al., 2008 [11]), as shown in Figure 1.3. In rats, the homology of structures with human prefrontal cortex is still debated, however the spatial distribution of cortical layers suggests homologies between rodents and primates (regarding granular areas, up to layer IV). Regarding OFC, neural activity and connectivity is largely shared between rats, primates and humans (Preuss, 1995 [12]). Despite a smaller size for OFC in rats, causing less ability to handle complex cognitive tasks as compared to primates, lesions in this area lead to the same dysfunction pattern across species in tasks with reversal learning and with reward devaluation (Stalnaker et al., 2015 [13]).

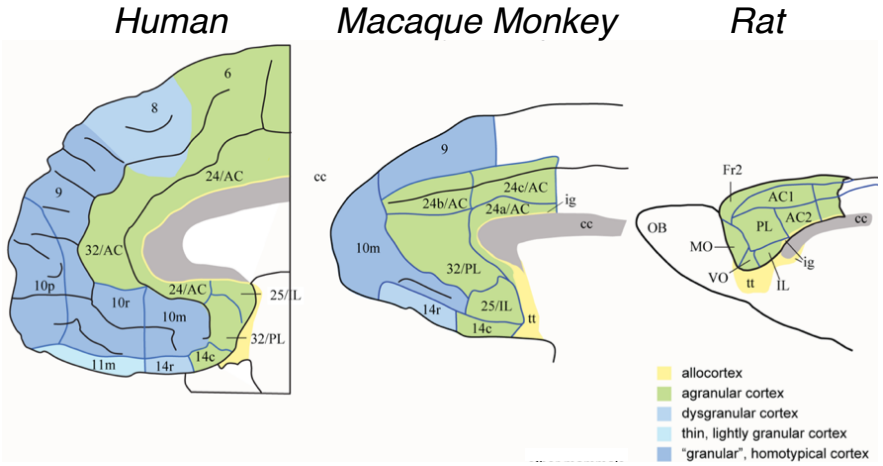


FIGURE 1.3: The human brain share homologies with other species (reproduced from Wise et al., 2008).

Medial area and cingulate sulcus share equivalent functional homologies between humans and macaque monkeys (Procyk et al., 2014 [14]). More precisely, the term anterior cingulate cortex (ACC) corresponds to different parts of the mid-cingulate sulcus according to different studies. Within the cingulate sulcus, we can distinguish the most anterior part (ACC), followed by the midcingulate cortex (MCC) also referred to as dorsal ACC (dACC). Furthermore, certain human subjects have an additional cingulate sulcus which is dorsal to the first one, named paracingulate sulcus (Petrides et al., 2012 [15]; Amiez et al., 2013 [16]).

In monkeys, the cingulate cortex presents similar cytoarchitectonic subparts, except for BA 32 which seems specific to humans (Vogt, 2009a [17]). Also, there is no paracingulate sulcus in monkeys. However, the most anterior part, corresponding to BA10, remains more developed in humans and seems to comprise cognitive processes that are specific to humans. Macaques contribution to neuroscience research involves local field potential (LFP) and unitary extracellular electrophysiology recordings, in behaving animal. However, this approach is not restricted to animals anymore. Recent studies start to use the same type of electrophysiology recordings in humans, with epileptic patients implanted with intracranial electrodes (Bonini, Burle et al., 2014 [18]).

1.1.4 Prefrontal cortex development and evolution during lifetime

In humans, prefrontal cortex is the last brain area to mature. Its development starts early in fetal life, in parallel with sensory and motor regions development, but is not over at birth and keeps growing during childhood and adolescence, up to 20 years old. The prefrontal endogenous circuits, driven by sensory stimulations, develop mostly during prenatal life, while the "cognitive" circuit appears at 7-12 months old. The maximal number of synapses and the complete maturation of certain cortical layers take place during the first few years of life (Gazzaniga, Chapter 2, 2009 [19]). At that moment, the number of synapses is much higher than in adults. The presence of extra-synapses allows to selectively stabilize certain functional circuits more than others, in response to various environmental stimuli and experiences, through intense pruning of supernumerary synapses. Synaptic connectivity exhibits initial exuberant production followed by gradual pruning (4-6 years old), with synapses density decreasing. The adult brain is then much less plastic.

Our faculty of judgment and decision is thus not complete until the prefrontal cortex is fully set up. Myelinization is not over until the second decade of life. Its development particularly depends on the amount and nature of exposure to stimuli, particularly to social stimuli, that are often complex and ambiguous. Blakemore's team has shown that prefrontal cortex in relation with social cognition keeps developing until late adolescence (Blakemore, 2010 [20]). These changes in behavior and cognitive skills are accompanied by changes in brain structure and in grey matter volume, regarding for example medial prefrontal cortex (Blakemore, 2008 [21]). Sense of self and relational reasoning also expand during adolescence (Dumontheil et al., 2010 [22]). Thus, executive function, which underlie our faculty of judgment and our sense of responsibility is not complete until the age of 18-20 years old. In the next section, we will now describe diseases arising as a consequence of PFC dysfunction.

1.1.5 Neuropsychiatric diseases involving prefrontal cortex dysfunction

Besides vascular lesions, many neuropsychiatric disorders in humans involve specific prefrontal deficits.

Obsessive-compulsive disorder. Patients with obsessive-compulsive disorder (OCD) display dysfunctional activity in orbitofrontal cortex, causing less behavioral flexibility (Chamberlain et al. 2008 [23]), as well as abnormal fronto-striatal loops functioning. OCD also involves basal ganglia dysfunction, leading to compulsive and repetitive behaviors (Baxter et al., 1992 [24]).

Addiction. Original addiction studies have focused on the reward circuit deficits in subcortical regions, such as ventral tegmental area. However, a growing body of evidence, coming from neuroimaging studies, revealed a key involvement of prefrontal cortex in drug addiction (Goldstein and Volkow, 2011 [25]), with an abnormal cognitive, motivational and emotional functions regulation. These studies indicated a decrease in cognitive control and in self-control in general, and a decrease of the ability to inhibit drives, characterized by a self-awareness lowering in intoxication periods (Baler and Volkow, 2006 [26]). Specifically, orbitofrontal and anterior cingulate cortices dysfunction implies over-saliency of stimuli related to addiction and under-saliency of other reinforcers.

Schizophrenia. In schizophrenia, post-mortem studies revealed a reduced brain volume, in particular in PFC and hippocampus, accompanied with abnormal cellular size, dendritic density and neural distribution. At the cellular and molecular levels, schizophrenic brain exhibits abnormal synaptic pruning during adolescence and early adulthood, corresponding to the symptoms onset. At the cognitive level, perceptual decision-making in schizophrenic patients is characterized by an over-dependence on prior expectations, despite sensory evidence being in contradiction with their prior expectations (Blackwood et al., 2001 [27]). This tendency to base decision on less evidence than healthy subjects has been termed "jump-to-conclusion" bias (Moritz et al., 2005 [28]). More precisely, Jardri and Deneve proposed a hierarchical neural network explaining circular belief propagation (Jardri and Deneve, 2013 [29]). This circular belief propagation results in abnormal interaction between top-down and bottom-up information (Fletcher and Frith, 2008 [30]). Their model explained schizophrenic patients' inflexible beliefs (Woodward et al., 2008 [31]) as well as their overconfidence in front of probabilistic choices. Moreover, the over-reliance on prior expectations hypothesis is supported by several data sets (Barbalat, Rouault et al., 2012 [32]; Chambon et al., 2011 [33]). Lastly, Barbalat and colleagues tested schizophrenic participants in a task involving top-down

cognitive control and maintenance of information from past events. Participants with schizophrenia had increased episodic control but had impaired contextual control (Barbalat et al., 2009 [34]). In addition, schizophrenic patients were impaired in effective connectivity within different lateral prefrontal cortex subparts, leading to a top-down control disconnection (Barbalat et al., 2011 [35]).

Historically, studying patients have provided some insight about the PFC functional roles. We will review in the next section the main theories of PFC functional architecture.

1.2 Functional and anatomical organization: main theories of prefrontal cortex function

In this section, we describe the proposed functions for the principal subregions of human prefrontal cortex. Roughly, human prefrontal cortex is organized around three main axes (Figure 1.4):

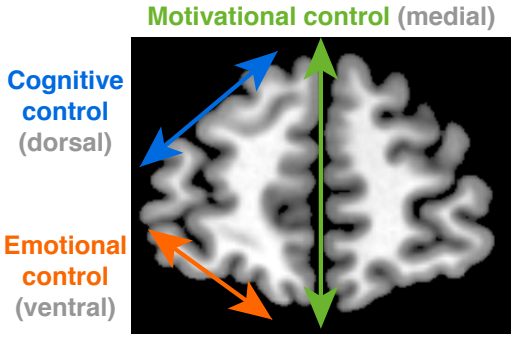


FIGURE 1.4: Prefrontal cortex and action control (coronal slice).

- **Motivational control**, which refers to drives, underlying voluntary action.
- **Cognitive control**, which refers to rules and choices.
- **Emotional control**, which refers to preferences.

Here, the term "control" refers to processes that are not *automatic* but *controlled* responses. First, we will see that dorsolateral prefrontal cortex (Figure 1.5) is responsible for top-down cognitive action control, while ventrolateral prefrontal cortex is related to motor inhibition and updating action plans. Next, we will see that ventromedial and orbitofrontal cortices (Figure 1.5) are heterogeneous brain regions, involved in particular in the outcomes and goods valuation, and in the values and emotions integration.

The following part will be dedicated to dorsomedial and cingulate cortices, implicated in motivation and performance monitoring. Finally, we will discuss the most influential accounts proposed to underlie frontopolar cortex, the most anterior part of the human brain.

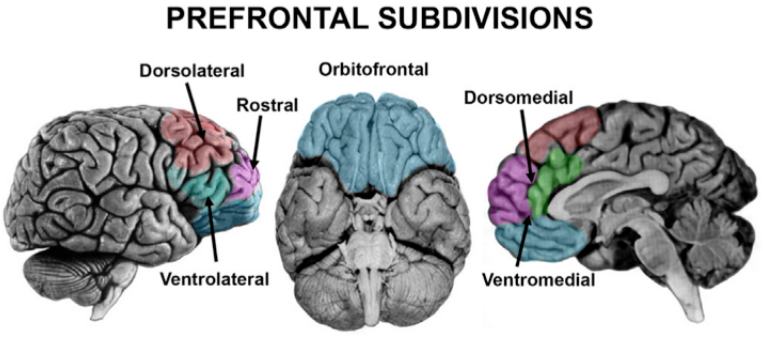


FIGURE 1.5: Main anatomical subdivisions within prefrontal cortex (reproduced from Szczepanski and Knight, 2014).

1.2.1 Lateral prefrontal cortex and hierarchical cognitive control

Lateral prefrontal cortex is implicated in **goal-directed behavior**. As such, it implements the behavioral adjustments that the medial PFC indicates, maintaining representations despite interference from distractors or irrelevant events until a goad is achieved. Lateral PFC is able to inhibit spontaneous responses before a motor action is executed. Lateral PFC is more activated following error trials, providing evidence for an increase in cognitive control for subsequent trials. As such, lateral PFC implements cognitive control adjustments.

Koechlin and colleagues have demonstrated a hierarchy in cognitive control within lateral prefrontal cortex, according to the information level. Here information is understood in the sense of Shannon information theory. At the sensory level, control is implemented in lateral premotor cortex, to select responses to stimuli (Koechlin et al., 2003 [36]). Certain neurons in lateral premotor encode planning an impeding movement and motor preparation. At the contextual level, caudal lateral PFC regions subserve control, in relation to external contextual cues associated with stimuli. Critically, contextual control is only engaged when current task contingencies require it (Collins and Frank, 2013 [37]). Finally, episodic control is implemented in rostral lateral PFC areas, given past behavioral episodes or internal goals, controlling more caudal regions in a "cascade" model. Top-down control is thus implemented according to a hierarchy in information processing and map onto a hierarchy in functional brain regions (Figure 1.6).

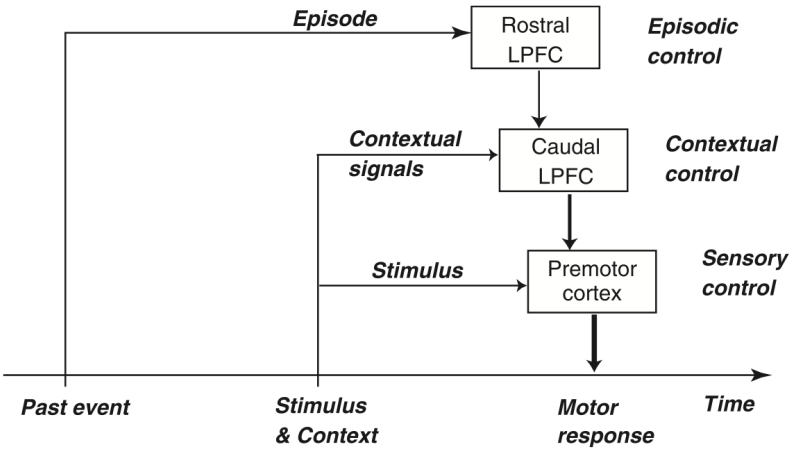


FIGURE 1.6: The cascade model of top-down cognitive control within lateral PFC (reproduced from Koechlin et al., 2003).

Rules implementation. In line with the notion of cognitive control, lateral PFC is involved in rule-based normative behavior. For example, Ruff and colleagues were able to increase or decrease compliance to social normative rules in humans, when manipulating right lateral PFC using transcranial direct current stimulation (tDCS) (Ruff et al., 2013 [38]). Behavioral rules neural substrates are found in lateral PFC in match-to-sample tasks. Ventrolateral PFC maintains rule representations to implement rule-based behavior (Sakai and Passingham, 2003 [39], 2006 [40]). Lateral PFC is not involved in encoding simple stimulus/reward rules but is necessary to encode more abstract high level rules and behavioral strategies (Bunge et al., 2005 [41]; Genovesio et al., 2005 [42]).

Working memory. The working memory concept refers to the ability to maintain relevant information to perform a task at hand, in the short-term, at a more abstract level than sensory information processing. This temporary information maintenance allows learning, comprehension and reasoning (Baddeley, 2010 [43]). Working memory is more about function than about contents. It includes an attentional focus mechanism, which consists of a bottleneck, meaning that only a limited amount of information can be handled at the same time (Oberauer, 2002 [44]; Oberauer and Kliegl, 2006 [45]). This mechanism allows to select sets or representations by determining priorities between various informations. Working memory function is critically dependent on lateral PFC (Levy and Goldman-Rakic, 1999 [46]), for instance for overcoming interfering stimuli.

In summary, lateral PFC is highly specialized regarding its anatomy and function. It has an integrative and adaptive role in a range of executive control behaviors, including retention, information manipulation and retrieval to achieve long-term goals, via action planning (Fuster, 2001 [47]), as well as response inhibition and rules implementation.

Ventrolateral PFC is involved in active information retrieval and selection, whereas dorsolateral PFC is rather involved in very controlled processes.

1.2.2 Ventromedial prefrontal cortex and orbitofrontal cortex

These two adjacent regions are sometimes similarly labelled in neuroimaging studies. Anatomically, they correspond to the two most ventral regions of medial PFC, associated with emotional/affective control (Figure 1.5).

Primarily, a large body of evidence supports the idea that ventromedial prefrontal cortex (vmPFC) and adjacent orbitofrontal cortex (OFC) encode "economic" value of goods or stimuli (Padoa-Schioppa and Assad, 2006 [48]; Lebreton et al., 2009 [49]; Prevost, Pessiglione et al., 2010 [50]; see Clithero and Rangel, 2013 for a review [51]). In neuroimaging experiments, vmPFC activity correlated with a "common currency" value for different types of goods (Chib et al., 2009 [52]) as well as with subjects' willingness to pay for food items (Plassmann et al., 2007 [53]).

However, we will see in the next chapter that reward *value* is a loosely defined concept (O'Doherty, 2014 [54]; Jessup and O'Doherty, 2014 [55]). Given the studies, it encompasses as far as reward identity, reward saliency, reward probability, etc.

Other pieces of evidence support a crucial role for vmPFC to make value-based *inferences* rather than simply retrieving values. For example, in a probabilistic reversal learning task under fMRI, vmPFC activity was found to be rather consistent with abstract hidden states inferences than with a reinforcement learning model (Hampton et al., 2006 [56]). Moreover, using neuronal recordings in rats, Jones and colleagues elegantly demonstrated that OFC is critically implicated to compute inferred values for decision (Jones et al., 2012 [57]). However, Roy and colleagues argued that vmPFC does not encode value per se, but encodes an "affective meaning" that is constructed from value. This affective meaning would be constructed from value by using other conceptual information, in order to give value its meaning in terms of behavior (Roy et al., 2012 [58]).

Despite a growing number of experimental studies implicating OFC in a large variety of computations (value, prediction errors, and their assignment to distinct causes: credit assignment (Walton, Behrens et al., 2010 [59], 2011 [60]), stimulus/outcome associations encoding ...), the exact role of OFC remains unclear (Stalnaker et al., 2015 [13]; Rudebeck and Murray, 2014 [61]). One of the most influential accounts to date for explaining OFC function across various datasets views OFC a "cognitive map of task space" (Wilson et al., 2014 [62]). OFC would encode the definition of a map of the current task-sets space, allowing for unlearning of old rules to set up new ones, and for

guiding behavior in the case of fictive learning (i.e. imagining outcomes that have never been encountered before, simulating possible outcomes, etc). It relies on the definition and position within a state space. Therefore, simple learning is still possible without OFC but as soon as the task requires more abstract inference, OFC remains necessary. Thus, OFC would acquire and maintain associative representations to guide behavior, in relation with hippocampus and striatum, to which it is connected.

We presented here briefly the main theories of ventromedial and medial orbitofrontal cortices, but the details and alternative theories regarding value processing in these regions will be presented in Chapter 2.

1.2.3 Dorsomedial prefrontal cortex and cingulate cortex

Dorsomedial prefrontal cortex (dmPFC) is a key node in the prefrontal network for decision-making and action control. In primates, it corresponds to the areas BA8m, BA9m and BA10m (medial parts). Its ventral limit corresponds to the named anterior cingulate cortex (ACC) and, more posterior, to the midcingulate cortex. The term anterior cingulate cortex (ACC) corresponds to different parts of the mid-cingulate sulcus according to different studies. Within the cingulate sulcus, we can distinguish the most anterior part (ACC), followed by the midcingulate cortex (MCC) also referred to as dorsal ACC (dACC) (Vogt et al., 2005 [63]; Procyk et al., 2014 [14]). These regions are found to be recruited in a huge number of situations, e.g. emotion processing, learning, motivation, error detection, reward processing, action/outcome evaluation and decision-making (Devinsky et al., 1995 [64]). We will discuss the main influential accounts of dorsomedial and cingulate functions.

Error Monitoring. dmPFC is implicated in error detection and subsequent behavioral adjustment. One of the main results refers to the *error-related negativity (ERN)*, an evoked potential that appears when the subject realized he/she made an error. The source in which ERN originates seems to be in dmPFC (Gehring et al., 1993 [65]; Holroyd et al., 2002 [66]). The ERN signal (Figure 1.7) is part of event-related brain potential that is generated when subjects made errors in psychophysics experiments. The ERN signal is thought to drive learning, although some people learn more from their errors while other people learn more from positive feedbacks (Frank et al., 2005 [67]). A more recent study using source localization precised the origin of ERN rather in premotor area/SMA, while the error-related positivity (ERP) was localized in ACC (caudal, BA24) (Herrmann et al., 2004 [68]).

However, this region might not selectively respond to errors but also monitors both correct and incorrect feedbacks. Specifically, Roger and colleagues showed that an anterior

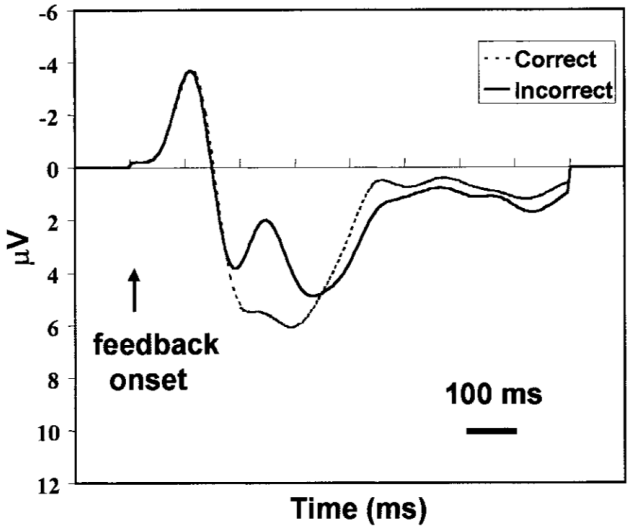


FIGURE 1.7: The Error-Related Negativity, elicited by the feedback apparition, is stronger for incorrect trials (reproduced from Holroyd et al., 2002).

part of the cingulate area was associated with correct outcomes, suggesting a common generator for correct and incorrect responses, with signal amplitude varying according to correctness (Roger et al., 2010 [69]).

A recent study also implicates SMA in error detection and online correction (Spieser et al., 2015 [70]). Using EMG, Spieser and colleagues were able to identify a role for SMA in inhibiting errors and correcting them before a motor response was provided. In addition, they were able to prevent impulsive errors using tDCS (Spieser et al., 2015 [70]).

Conflict theory. When and how is cognitive control recruited in lateral PFC? In a series of papers, Botvinick and colleagues proposed that ACC activity increases when *conflict* between competing responses arises i.e. when there is a conflict in information processing, thus generating the need for more cognitive control (Botvinick et al., 2001 [71], 2004 [72]; Shenhav et al., 2014 [73]). More precisely, conflict may arise when the subject has to override a predetermined response, or when visual stimulus and motor response directions are incongruent. Flanker task and Stroop effect are examples of experimental set-ups causing conflict; even if eventually no error is made. Conflict also arises when responses are underdetermined (many possibilities), thus generating higher activity in ACC. Thus, ACC is recruited in tasks where there is a high demand for cognitive control, the level of which would be regulated by an interaction medial/lateral PFC. A meta-analysis of datasets identifying conflict (Barch et al., 2001 [74]) revealed that not only ACC but mainly MCC and sometimes even dmPFC were recruited when the task at hand triggered conflict (Figure 1.8).

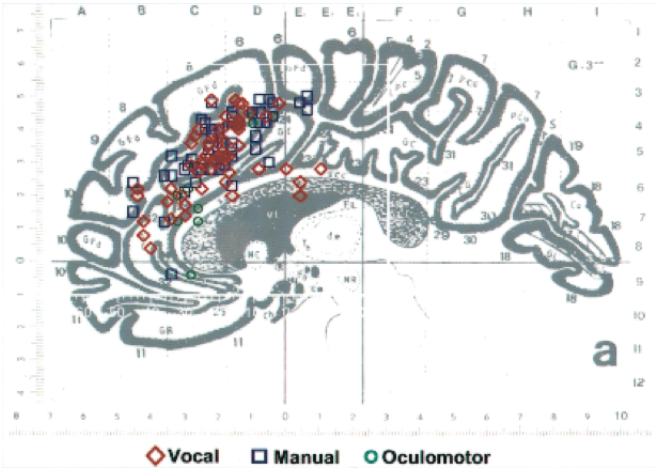


FIGURE 1.8: Plot of cingulate activations related to conflict with various response modalities, from a literature review (reproduced from Barch et al., 2001).

However, many datasets have now put into question conflict theory. In particular, Burle and colleagues showed that competition between responses was not required for interference to happen. The ERN signal duration was related to the time necessary to "correct" partial errors arising, lasting up to the moment that the error was "corrected" (Burle et al., 2008 [75]). Moreover, Burle and colleagues showed that the engagement of executive control was not directly related to the amount of conflict, as measured by electromyographic recordings (Burle et al., 2005 [76]).

Motivational control. In line with the role of dmPFC and ACC in regulating cognitive control engagement, it has been proposed that dorsal regions of the cingulate cortex are responsible for the motivation for action and the notion of "wanting" something, regulating the level of lateral PFC subparts recruitment. Varying monetary incentives via visual cues, Kouneiher and colleagues demonstrated two motivational control levels within medial PFC, mapping onto cognitive control levels within lateral PFC and "energizing" them (Kouneiher et al., 2009 [77]). More precisely, activity in pre-SMA was associated with contextual motivation, whereas episodic motivation triggered dACC activity.

Mentalizing. dmPFC is a key node in the mentalizing network and is implicated in social cognition (Eickhoff et al., 2014 [78]). Mentalizing refers to the capacity to understand the mind of others, for example by maintaining representation of their preferences (Kang et al., 2013 [79]). Mentalizing also relies on temporo-parietal junction (TPJ) and posterior cingulate cortex (PCC). In a task designed to trigger altruistic behavior, Waytz and colleagues found that dmPFC activity predicted both monetary donations to other people and time dedicated to help others (Waytz et al., 2012 [80]).

Exploration and Foraging. Should I stay or should I go? Ecological decision-making comprehends a trade-off between sticking into the same environment (exploitation) or looking for new choice options (exploration). Exploring critically engages ACC. In macaque monkeys, neurons within dACC encode relative evidence in favor of foraging i.e. switching to another source of potential rewards (Hayden et al., 2011 [81]). These neurons fired at each decision to stay, up to a certain threshold from which the animal switched (Figure 1.9).

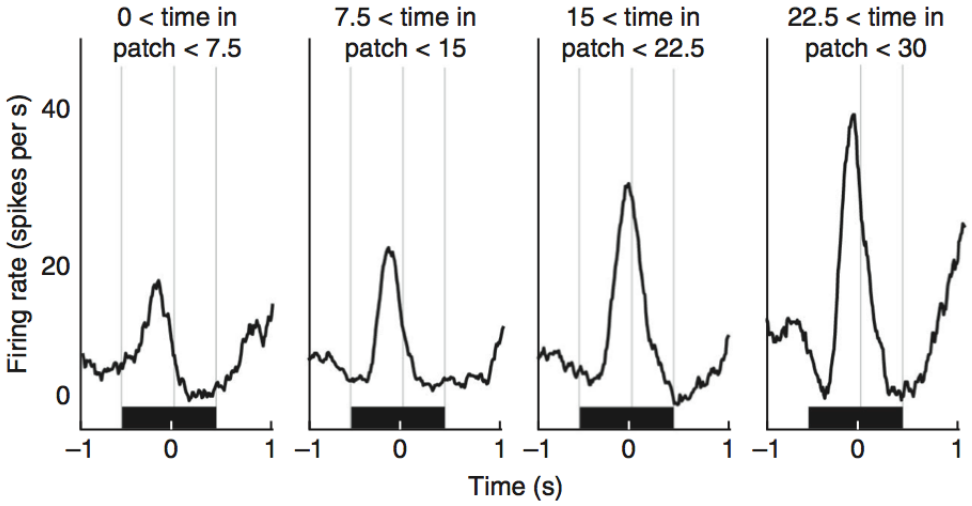


FIGURE 1.9: The firing rate in dACC neurons increased with time spent in a food patch, up to a certain threshold triggering patch leaving and exploration (reproduced from Hayden et al., 2011).

The threshold limit for foraging was dependent on travel time, modeling the necessary time to move to a new option. This travel time represents a "cost" associated with foraging. In humans, in a task involving evidence seeking to make reward-based decisions, anterior cingulate activity increased when subjects sampled more evidence as compared to when they engaged in a decision (Furl and Averbeck, 2011 [82]). Similarly, Kolling and colleagues designed a fMRI paradigm in which human subjects had to trade-off exploiting a current set of choice options, or exploring a new environment with a new set of choice options, but sustaining a cost for this foraging (Kolling et al., 2012 [83]). They revealed that the ACC/dACC encoded the foraging cost, as well as the foraging environment mean value. However, certain authors have contested this result, arguing that dACC rather represented choice difficulty, which was confounded with foraging value (Shenhav et al., 2014 [73]).

Thus, exploration implicates a notion of cost. Leaving a current patch of resources for another location is accompanied with some uncertainty and with some cost (e.g. travel

cost). This aspect relates to the conflict, i.e. an increased cognitive cost when there is conflictive information to process.

Action-outcome predictor. It could be that the ACC and dmPFC are not specifically involved in error detection but more generally encodes outcomes or stimulus/outcome links. Indeed, Matsumoto and colleagues recorded neurons in MCC that fire for specific action-outcome combinations, evoking an action plan or selecting actions among several possibilities in relation to a goal (Matsumoto et al., 2003 [84]). A theoretical work that could unify the variety of previous findings regarding medial PFC, especially gathering the error monitoring and conflict accounts, proposes that the medial PFC acts as an action/outcome predictor, evaluating the discrepancy between predicted and obtained outcomes, irrespective of outcome valence (Alexander and Brown, 2011 [85]). The medial PFC is viewed as an action/outcome predictor, detecting mismatch between predicted and real outcome. The proposed model represents multiple action/outcome associations at the same time. It then measures the *surprise* generated by the feedback, which reflects the discrepancy between actual and observed outcome, regardless of outcome value. Thus, the medial PFC would signal the unexpected non-occurrence of predicted events. The model is able to account for and reinterpret a range of experimental findings regarding ACC and MCC function, such as error monitoring, conflict, action values prediction, etc. (Silvetti et al., 2013 [86]).

In summary, the dACC/MCC is engaged when *switching away* from the current course of action or from a default behavior (Boorman et al., 2013 [87]).

1.2.4 Frontopolar cortex

Frontopolar cortex, sometimes called frontal pole (BA 10), corresponds to the most anterior part of the brain and is phylogenetically the most recent part. It is associated with high level control. Patients with lesions in this area usually perform poorly in open-ended or novel environments or in environments with unusual structure or attentional demand, while being able to carry out normally tasks that are supposedly requiring general PFC function. We review here the main theories examining frontopolar cortex function.

The gateway hypothesis: switch between "in" and "out" modes. An attentional theory suggest that BA 10 arbitrates between an internal mode ("stimulus-independent"), in which the subject is focused on her own thoughts and intentions, and an external mode ("stimulus-oriented"), in which she responds to environmental stimuli (Burgess et al., 2007 [88]). Indeed, this area is required when the subject is attending to her own mental states (Frith and Frith, 2003 [89]). According to this theory, frontopolar

cortex might not support costly complex cognitive computations, as opposed to other subparts of prefrontal cortex, but arbitrates between different attentional modes.

Monitoring of alternative courses of action. Frontopolar cortex is involved in "branching" control, meaning that it maintains information about a pending task that is interrupted while another task is performed (Hyafil and Koechlin, 2007 [90]). The arbitration between dedicating cognitive resources to a current task and retrieving a pending task is based on future expected reward associated with each task. This branching control enables to put aside a current task while performing another task and to go back to it afterwards, hence allowing for multitasking (Koechlin et al., 1999 [91]). This function enables humans to maintain long-term goals while being able to respond to immediate stimuli or environmental demands.

Frontopolar cortex was also shown to be involved in exploration of alternative actions (Daw et al., 2006 [92]). In Daw and colleagues' experiment, subjects had to choose between four bandits providing stochastic rewards, with the average reward for each bandit continuously drifting across the experiment, thus triggering the need for constant exploration of alternative bandits. Daw and colleagues found the frontopolar cortex to be recruited selectively for exploratory trials, corresponding to trials in which subjects sampled a different bandit that the one they thought had the highest expected value. Consistently, Boorman and colleagues found that frontopolar cortex tracked the relative advantage of the alternative option, in a probabilistic learning task (Boorman et al., 2009 [93]). This theory is further supported by recent results regarding the monitoring of alternative task-sets reliabilities in frontopolar cortex, in a task involving learning, creating and adjusting behavioral task-sets (Donoso et al., 2014 [94]).

Metacognitive evaluation. Gray matter volume in the frontal pole has been shown to correlate with metacognitive abilities, i.e. the capacity to evaluate one's own performance (Fleming et al., 2010 [95]). In a perceptual decision-making task followed by confidence ratings, Fleming and colleagues revealed that across individuals, the better the introspective accuracy, irrespective of objective performance, the larger the gray matter volume in anterior PFC. However, the reliability of these metacognitive judgments also involves dorsolateral PFC and cingulate cortex (Fleming et al., 2012 [96]).

1.2.5 Conclusion

We have reviewed in this first chapter the most influential theories explaining prefrontal cortex functional architecture. We have highlighted that the critical feature of prefrontal executive control lies in its evaluation function: error monitoring, action outcome monitoring, alternative courses of action monitoring. These monitoring processes (medial

PFC up to frontopolar) confer to prefrontal cortex its ability to subsequently adjust immediate and future action (lateral PFC). The monitoring function that PFC subserves is at the core of human adaptability, allowing us to behave flexibly, not only reacting to external stimuli but acting in relation to internal mental states. Accordingly, executive control relies on evaluating action outcomes to adjust immediate and future action. However, action outcomes may convey several types of value signals. In the next chapter, we will focus on the affective value of action outcomes.

Chapter 2

Affective values in human decision-making

Action outcomes can convey a positive or a negative value, in the form of rewards and punishments. Therefore, action outcomes transfer an affective value that is going to influence choices in return. Here, *affective value* is understood as reward amplitude, which we could also have named "rewarding value". Unlike its common meaning, the term *affective* here refers to the motivational properties of outcomes for action, rather than the emotional properties. In this chapter, we will focus on the notion of affective value of rewards and how it drives learning and decision-making, from psychological, theoretical and cerebral points of view.

2.1 Affective values: psychological and theoretical aspects

2.1.1 The notion of affective value, based on rewards and punishments

The concept of affective value is a behaviorally relevant one. It is thought to generate motivation and to drive action. Broadly, animals seek positive rewards (O'Doherty, 2014 [54]). It also relates to the notion of pleasure and pleasantness that we experience, anticipate or even imagine (Berridge and Robinson, 2003 [97]). This could be measured by agreeableness or desirability subjective reports. Importantly, affective values are not binary. They can vary parametrically and continuously. In behavioral economics, value refers to the notion of utility (expected utility/experienced utility). Economic value corresponds to the quantity that an agent tends to maximize. It is used to examine consumer behavior. Finally, the concept of value is at the core of subjective preferences,

that can be elicited with binary choices, although expression of subjective preferences often depart from rationality (Kahneman and Tversky, 1979 [98]).

We can distinguish primary rewards (e.g. food, sex) that have a physiological meaning from secondary rewards that convey a more abstract value signal, for example money. However, it remains difficult to separate pure value from components that support the construction of a value signal upstream.

2.1.2 Expected utility theory and prospect theory

Facing a choice between two items, subjects compute a **subjective value**, via a valuation process, and choose the highest of both values. Choice psychology often observed departures from rationality (Tversky and Thaler, 1990 [99]). Facing twice the same choice, subjects do not always choose the same item. Despite detected inconsistency, it seems hard for participants to resolve it, even if they aim at showing internal coherence (Tversky and Kahneman, 1981 [100]). This internal inconsistency, namely, cognitive dissonance, leads participants to try to match their choices so as to respect internal consistency with themselves, otherwise resulting in psychological discomfort (Festinger, 1962 [101]; Izuma et al., 2010 [102]; Salti et al., 2014 [103]). Expected utility theory states that subjects aim maximizing expected utility, which corresponds to maximizing subjective value and subsequent satisfaction obtained from goods or rewards.

Several mathematical functions have been used to describe choice between items of similar expected utility. A first possibility would be to systematically choose the item with the highest expected utility: the maximum among all subjective affective values. However, this possibility does not account for the fact that subjects sometimes choose the lowest of two options, in order to explore alternative options. Another possibility is to introduce stochastic choice with the *softmax* function (Figure 2.1). The softmax function is a way to model the choice probability according to the subjective value of two items in a binary choice setting. One item or action is stochastically selected according to the difference between each item's expected utility (Luce, 1977 [104]). The inverse temperature parameter regulates the sigmoid slope, and the amount of exploratory choices. Exploratory choices correspond to the proportion of choices in which the lowest valued of two actions is occasionally preferred. A large inverse temperature corresponds to almost deterministic choices, whereas a small inverse temperature corresponds to more noisy, and at the extreme, more random choices.

Prospect theory adds a few concepts to expected utility theory, to explain apparent biases and inconsistencies in choice, especially in risky prospects. In particular, prospect theory states that choice is dependent on:

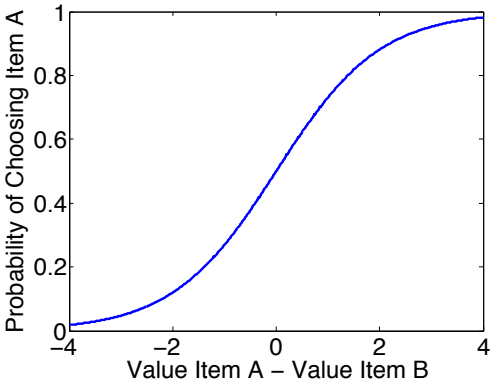


FIGURE 2.1: The softmax function is a way to model the choice probability according to the subjective value of two items in a binary choice setting (arbitrary value units).

- **Framing effects.** Subjects' response is dependent on the problem formulation. Two problems that are identical in terms of expected utility will lead to significantly different choices according to whether they are framed as gains or losses. Other authors have identified that choices are also dependent on the visual presentation of the gains at stake in experimental settings, for example whether money is depicted as digits or as piles of coins (Sharp et al., 2012 [105]).
- **Loss aversion.** The loss of a particular amount is more aversive than the gain of the same amount is rewarding (Tom et al., 2007 [106]).
- **Probability distortions.** Subjects tend to overestimate the probability of very unlikely events, while underestimating the probability of very certain events, resulting in an inverse sigmoid distortion of probabilities representation.

In our fMRI study, we tested the prospect theory model, which is very general, to explain participants' choices. Eventually, we found that it fitted less parsimoniously the behavioral data than other alternative models (see Chapters 7 and 8). So, we have seen that economic value of goods or items, understood here as "affective" value, is driving choice, as expressed according to subjective preferences. We will describe in the next section the psychological theories of how these values are driving our behavior.

2.1.3 Rewards are driving learning: pavlovian and instrumental conditioning

Conditioning theories were first studied in animals. If an action triggers a positive outcome as a consequence, the animal will tend to repeat that action. On the contrary, if an action leads to a negative outcome, the agent will tend to avoid it in the future.

Thus, action outcomes drive learning through affective value signals. Positive outcomes will generally elicit subsequent approach behavior, whereas negative outcomes will elicit subsequent avoidance behavior.

Historically, **Pavlov** (1849-1936) observed that if a neutral irrelevant stimulus was paired with a behaviorally meaningful stimulus that elicited a response (unconditioned or 'reflex' response), after a number of repetitions, the neutral stimulus alone was sufficient to elicit the response. This phenomenon was termed *classical (or pavlovian) conditioning.* The concomitance of the association between the neutral (conditioned) and the behaviorally meaningful (unconditioned) stimuli elicits response learning (Figure 2.2).

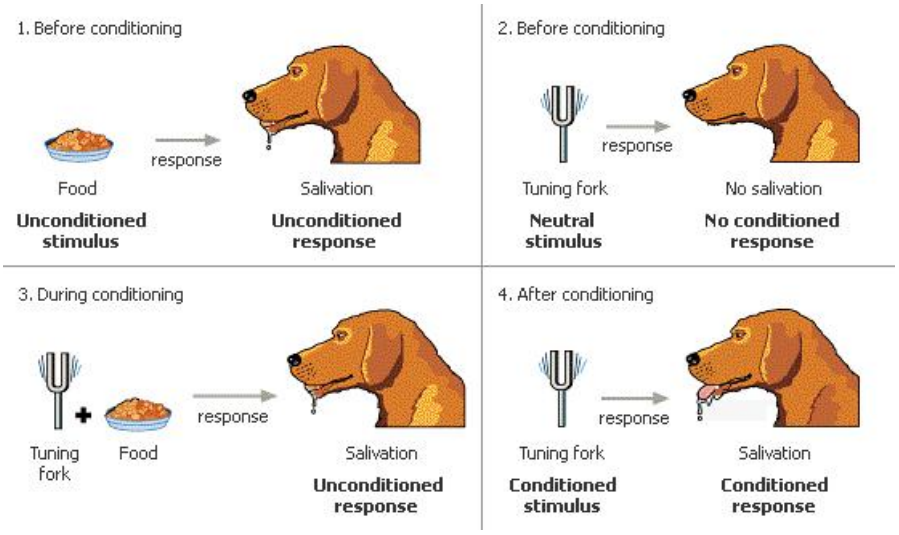


FIGURE 2.2: Pavlovian conditioning (source: http://schoolworkhelper.net/).

A few decades later, **Thorndike** (1874-1949) law of effect described the idea that a positive outcome will increase the probability of that action to occur again in the same situation. Unlike his predecessors, Thorndike believed that before the first reinforcer, the chosen actions were random, until the moment that, by chance, the animal finds a relevant action. After a number of repetitions, the animal was able to identify faster the relevant action to perform. Building on this work, **Skinner** (1904-1990) and the wave of behaviorists described *operant (or instrumental) conditioning.* Operant conditioning is defined as strengthening or weakening of association between a stimulus and an action given reinforcement obtained. As opposed to classical pavlovian conditioning, instrumental conditioning implicates the production of an action by the agent, not juste passive association between external stimuli. Moreover, Skinner provided detailed quantitative analysis regarding the time or the number of trials and errors that an animal executed before an appropriate behavior was learnt. Although he did not ignore the

influence of internal variables that could drive learning but are inaccessible to observation, he focused on the link between environmental stimuli and external, measurable behavior, without investigating the underlying internal mental representations.

Conditioning can generalize across other stimuli that share features with the neutral conditioned stimuli. The conditioned response can thus extend for example to stimuli with similar sensory properties. Nevertheless, the conditioned response appears in a specific situation and its generalization is limited. Another interesting trait of conditioning is its capacity of extinction. If the conditioned stimulus is presented a lot without the unconditioned stimulus that originally elicited a response, then the conditioned response might disappear. Thus, conditioning is reversible. But the reinforcer effect can be persistent. For example, subjects with addiction keep being sensitive to reinforcers, even if they report them to be no longer "valuable" rewards. Given the animal's behavior after extinction, we can distinguish **goal-directed** vs. **habitual** behavior. Usually, habits arise after a long temporal sequence with many repetitions of the behavior. In habits, it is the association between conditioned stimulus and response which mainly drives action. Consequently, after extinction, the animal would keep reproducing the actions that led to reinforcement. By contrast, in goal-directed behavior, it is the outcome affective value that drives action. Consequently, after extinction, the animal would gradually stop the actions that previously led to reinforcement, since they are no longer followed by a rewarding outcome.

Today, behaviorism have contributed to the cognitive behavioral therapies development. These therapies aim at dealing with the observable behavioral symptoms rather than focusing on internal mental states that might generate the symptoms. By manipulating the conditioning between stimuli and responses through reinforcement or extinction, these therapies have proven efficient to address adaptive problems such as anxiety or phobias.

2.1.4 Reinforcement learning computational models

The conditioning and behaviorism psychology has been formalized mathematically with reinforcement learning theories. This set of learning algorithms originally came from the machine learning field (Bishop, 2006 [107]; Sutton and Barto, 1998 [108]). Computational models of reinforcement learning provide a normative framework of how an agent can learn action affective value by interacting with the environment. Computational models of reinforcement learning are based on the concept of **prediction error**, which measures the discrepancy between the expected and the actual action outcome value (Rescorla and Wagner, 1972 [109]). As stated in Rescorla rule, the action affective value

V_t that led to an outcome r_t is updated according to:

$$V_{t+1} = V_t + \alpha(r_t - V_t), \tag{2.1}$$

in which α is the learning rate, modulating the degree to which the prediction error $r_t - V_t$ affects the chosen action value V_t. If α is high, recent outcomes matter more. At the extreme, if $\alpha = 1$, the action expected value V_t reduced to the last outcome value, r_t. If α is low, recent outcomes little modify the chosen action expected value V_t, and therefore a larger reward history for estimating that action value is taken into account. Thus, the agent learns the value V_t by experience, sampling from the environment through trial and error. Reinforcement learning as formulated above consists of a trial by trial continuous update, not sensitive to temporal blocks within learning or to the possible higher-order structures of the environment in which learning occurs. It is referred to as **model-free reinforcement learning**.

Sutton and Barto elaborate on this to allow information diffusion across contiguous time points, in the form of the temporal difference learning algorithm. Temporal difference algorithm includes a temporal discounting parameter that model the reinforcer devaluation with time (Sutton and Barto, 1998 [108]). In simple words, according to this algorithm, predictions are tuned to formulate more precise predictions about the future, with the discounting parameter modulating the impact of rewards across different time points. Yet, if the next state $t+1$ does not depend on the chosen action at t, the discount factor is unnecessary and the prediction error can be written as above. Another form of reinforcement learning algorithms is Q-learning, in which the value V_t corresponds to a state-action value and not only to an action value as in temporal difference learning. Its extension, the SARSA algorithm, is similar but differs regarding the control strategy. While Q-learning assumes an optimal policy for action selection at the next state and subsequent action values update, in SARSA the value of the actual chosen action is used for updating.

As opposed to behaviorism that only focuses on observable behavior (actions), reinforcement learning includes the notion of internal hidden variables, i.e. action value or state/action values, that shape observable behavior.

2.1.5 Model-based and model-free reinforcement learning

Model-based reinforcement learning includes a notion of internal state, that will be further developed in the next chapter. So far, the learning policies described above are comprised in what is called model-free reinforcement learning. It is based on learning "cached values" of the environment by trial and error, without any prior assumptions

about the environment structure. By contrast, model-based reinforcement learning includes a state model of the environment structure ("tree search"), which is an explicit representation of the world, on which learning is based. The agent learns on the basis of this internal states representation, without the need to sample every possible action, as opposed to model free RL. In practice, model based RL can rapidly becomes intractable, because of the huge number of possible states. But solutions have been proposed for example, pruning a number of states (a "branch of the tree").

Model-based and model-free RL differ according to their sensitivity to reward devaluation (Daw et al., 2005 [110]). In model-based RL, the outcome affective value back-propagates to all actions that have led to that particular terminal state where an outcome was obtained. In model-free RL, reward devaluation only affects the choice of the action that was the most proximal to the outcome, independently of the other states crossed beforehand. In that case, the reward devaluation effects will be slower. In a number of situations, the model-free system is faster and more efficient.

However, having the two systems in parallel is advantageous. It enables trading-off the habitual model-free system inflexibility, and the model-based system, more flexible but associated with a higher computational cost (Daw et al., 2005 [110]). Each system can be used in circumstances in which it is the most accurate. The arbitration between model-based and model-free reinforcement learning systems remains controversial (Dezfouli et al., 2013 [111]). Some have proposed that arbitration between both systems relies on their respective uncertainties, on each trial (Daw et al., 2005 [110]). Other authors have suggested the existence of a "responsibility signal" associated with each task-set driving behavior. A task-set is defined by the representation of a stimulus/action/outcome mapping. Choice is then controlled by the weighted average of the responsibility signals (Doya et al., 2002 [112]; Samejima and Doya, 2007 [113]). However, the latter proposal implies a task-set selection at each trial, whereas humans rather tend to adopt a default behavior and switch to exploration only when necessary. Behavioral (Figure 2.3) and neural hallmarks of both systems have been identified in the brain, implicating ventral striatum and lateral prefrontal cortex (Glascher et al., 2010 [114] ; Daw et al., 2011 [115]).

Conclusion. We have reviewed the main psychological observations and computational theories underlying affective values processing. Pavlovian and instrumental conditioning revealed how affective values conveyed by rewards and punishments shape learning and subsequent choices, in animals and in humans. We have then seen that mathematical models of reinforcement learning provide a normative framework for describing the computations supporting learning from affective values, centered on the notion of prediction

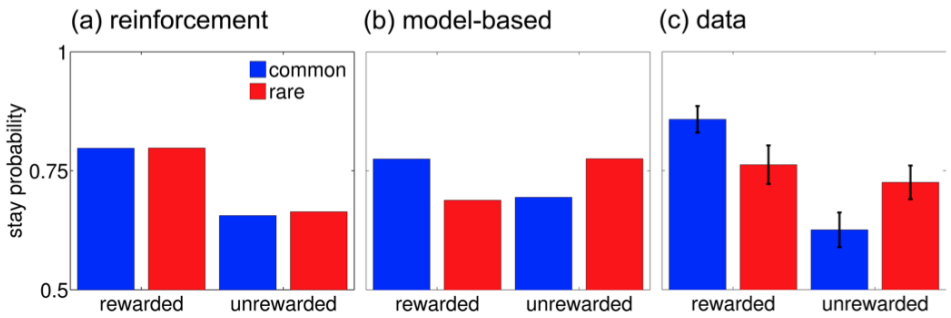


FIGURE 2.3: Subjects' behavior appears to be in-between model-based and model-free reinforcement learning predictions (reproduced from Daw et al., 2011).

error. In the next section, we will examine how these learning mechanisms based on affective values are implemented in the brain.

2.2 Affective values: cerebral aspects

How are affective values represented in the brain? Several cortical and subcortical areas are implicated in rewards affective value processing.

2.2.1 Basal ganglia

2.2.1.1 Subcortical basal ganglia anatomy

Basal ganglia refer to a set of subcortical nuclei (Figure 2.4). It comprises dorsal and ventral striatum (putamen and caudate nucleus), substantia nigra pars compacta (SNc) and pars reticulata (SNr), ventral tegmental area (VTA), internal and external globus pallidus (GPi, GPe), thalamus, hypothalamus and subthalamic nuclei.

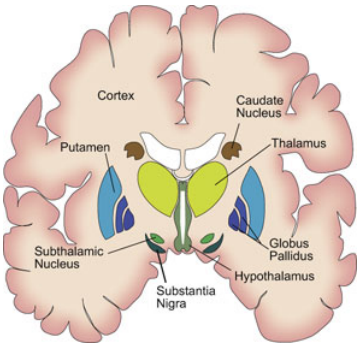


FIGURE 2.4: Basal ganglia anatomy (coronal slice) (reproduced from Adam, 2013).

These structures are connected to the neocortex via fronto-striatal loops (Haber, 2003 [116]) and are present in many species. Basal ganglia are involved in motor control and reward learning, through three main neurotransmitters projections (γ-aminobutyric acid (GABA), dopamine and glutamic acid). Dysfunction of dopamine direct or indirect pathways can lead to movement control disruption, for example in Tourette syndrome, Parkinson and Huntington diseases.

2.2.1.2 Electrophysiology and pharmacology studies show reward prediction error in dopamine neurons

Reward prediction errors representations were identified in dopamine neurons in the basal ganglia, using electrophysiological recordings in primates (Schultz et al., 1997 [117], 1998 [118]). Schultz and colleagues described a population of neurons that fire more with reward unexpected occurrence, less with reward unexpected non-occurrence, as compared to a baseline firing rate corresponding to reward expected occurrence (Figure 2.5).

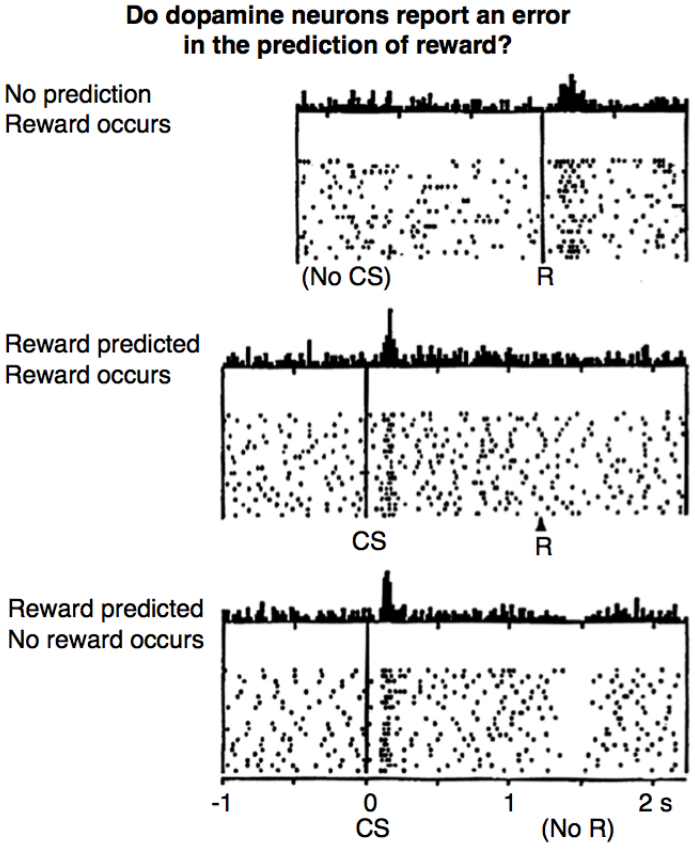


FIGURE 2.5: Midbrain dopamine neurons encode reward prediction errors (reproduced from Schultz et al., 1997).

Such a pattern is interpreted as coding the reward prediction error $r_t - V_t$ as described in the previous section with reinforcement learning algorithms. Parametric activity of these dopamine neurons modulates cortical regions, which in turn integrate prediction errors to form future predictions. In addition to prediction error signal, Fiorillo and colleagues identified a neural response sensitive to the amount of uncertainty, also in dopamine neurons (Fiorillo et al., 2003 [119]).

Further evidence comes from Parkinson patients studies with pharmacological dopamine manipulation. Patients with Parkinson disease usually receive medication to enhance the dopaminergic system, either in the form of dopamine agonists or in the form of levodopa which will be transformed into dopamine in the brain. Although these drugs generally improve motor symptoms, they convey side-effects regarding executive control and learning from positive and negative feedbacks. Frank and colleagues showed that Parkinson patients receiving levodopa were better at learning from rewards than from punishments, probably through reinforcing direct pathway (Frank et al., 2004 [120]). By contrast, without levodopa, Parkinson patients were better at learning from punishments. Therefore, the authors demonstrate that dopamine level modulates learning, from both positive and negative outcomes, in a dynamic way. Palminteri and colleagues replicated and extended these results in the case of subliminal learning (cues are not consciously perceived) and in the case of patients with Tourette syndrome, which present an inverse pattern regarding dopamine and motor deficits as compared to Parkinson patients (Palminteri et al., 2009 [121]).

2.2.1.3 The contribution of neuroimaging studies

Evidence from non-human primate electrophysiology and pharmacological manipulations in humans was strengthened by neuroimaging data in humans. Combining primary rewards (erotic images) and secondary, more abstract rewards (amounts of money), Sescousse and colleagues revealed a common network for affective values processing, implicating the ventral striatum and midbrain, as well as other cortical regions (Sescousse et al., 2010 [122]).

Further imaging results came from the first model-based fMRI studies. For example, O'Doherty and colleagues scanned human subjects while performing a Pavlovian and an instrumental task to obtain juice reward. Reward prediction errors neural correlates were found in ventral striatum, comprising nucleus accumbens and ventral putamen (O'Doherty et al., 2004 [123]). Using a reinforcement learning algorithm, they found that the dorsal striatum was engaged only with instrumental conditioning.

Model-based fMRI. Essentially, this method allows to identify regions that specifically correlate with a model's variable (O'Doherty et al., 2007 [124]). Rather than solely identifying locations, model-based fMRI informs about the cerebral implementation of the cognitive mechanisms that a computational model describes. The trial-to-trial variables are extracted from a computational model. Then, the variables are regressed against BOLD signal to identify voxels in which brain activity significantly correlates with each of the variables. We will be using the model-based fMRI approach in this thesis.

Going back to the basal ganglia contribution in representing affective values, the dopamine effects on learning from rewards and punishments via the basal ganglia have been tested in healthy human subjects. More precisely, Pessiglione and colleagues investigated the behavioral effects of two drugs modulating dopamine, enhancing dopamine production and a dopamine antagonist (Pessiglione et al., 2006 [125]). Using model-based fMRI in a learning paradigm, the authors reveal that the drugs differentially modulated the prediction error amplitude in ventral striatum. They excluded an effect of drugs on general mood or reaction times. Consequently, they observed at the behavioral level that subjects treated with dopamine enhancer better learn to choose to obtain positive outcomes, as compared with subjects treated with dopamine antagonist. However, regarding negative outcomes avoidance, there was no drug-induced modulation.

2.2.2 Medial prefrontal cortex

Monkey neurophysiology and human fMRI data support the affective values encoding primarily in vmPFC and adjacent medial OFC. A vast body of evidence, from both electrophysiology and neuroimaging studies, supports the idea that ventromedial prefrontal cortex (vmPFC) and adjacent medial OFC encode "economic" value of goods or stimuli (Padoa-Schioppa and Assad, 2006 [48]; Lebreton et al., 2009 [49]; Prevost, Pessiglione et al., 2010 [50]; see Clithero and Rangel, 2013 for a review [51]).

Moreover, Padoa-Schioppa and Assad recorded neurons in OFC that encode the economic value of juices that the monkey chooses to consume, irrespective of the action performed to receive them (Padoa-Schioppa and Assad, 2006 [48]). Similarly, Tremblay and Schultz recorded neurons in central OFC responding to the relative value of a juice, independently of the juice actual sensory properties (Tremblay and Schultz, 1999 [126]). Converging evidence in animal thus revealed that medial OFC appears to encode rewards affective values.

In humans, in a task involving ratings of stimuli affective value, followed by choices between stimuli to generate expression of subjective preferences, Lebreton and colleagues isolated a "brain valuation system" comprising ventromedial prefrontal cortex, ventral

striatum, posterior cingulate and hippocampus (Figure 2.6, Lebreton et al., 2009 [49]). Critically, the valuation system was active even when valuation of stimuli was irrelevant for the task at hand.

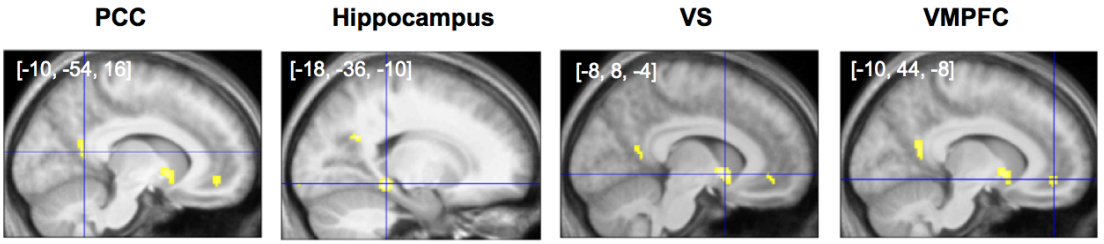


FIGURE 2.6: The brain valuation system encoding value assigned to images (reproduced from Lebreton et al., 2009).

This network, in particular the vmPFC, was found in several other studies involving valuation (Boorman et al., 2009 [93]; Clithero and Rangel, 2013 [51], Bartra et al., 2013 [127]). For example, Plassmann and colleagues designed an economic task evaluating subjects' willingness to pay for food items. To evaluate their willingness to pay, the authors used a Becker-DeGroot-Marshak auction, an economic procedure established to reveal the economic value that subjects attribute to items. They identified that vmPFC encoded subjects' willingness to pay for food items. Here economic value is understood in the sense of affective value, in terms of subjective preferences. Similarly, Chib and colleagues found that vmPFC encoded a value signal on a common scale regardless of the nature of the item that was evaluated (Chib et al., 2009 [52]), supporting the idea that the brain valuation system operates across domains [49]. Medial PFC is also implicated in the construction of a value signal from previously estimated values, along with the hippocampus, supporting the possibility of imagining outcomes (Barron et al., 2013 [128]).

More precisely, other authors have attempted to precise the exact nature of the value computations performed in vmPFC and adjacent OFC. Bouret and Richmond recorded neurons in ventromedial prefrontal and orbitofrontal cortex of behaving monkeys (Bouret and Richmond, 2010 [129]). They found that neurons in both regions encoded the value of task events. However, they revealed that vmPFC was crucial for internally driven values (e.g. self-initiated action), whereas OFC was critical for externally driven values, based on presented visual cues. Using neuronal recordings in rats, Jones and colleagues elegantly demonstrated that the OFC is crucially involved when inferring values for decision (Jones et al., 2012 [57]). In addition, Roy and colleagues argued that the vmPFC does not encode value per se, but encodes an "affective meaning" that is constructed from value. This affective meaning would be built from "pure" value by using other

conceptual information, in order to give value its meaning in terms of behavior (Roy et al., 2012 [58]).

vmPFC/medial OFC thus appears to encode a general affective value signal, whether it be a "decision value", "feedback experienced value", "goal value" or "anticipated value". vmPFC critically represents an option value when a decision is made to engage with this option (Kolling et al., 2012 [83]). However, we will see in the next section that other pieces of evidence rather support a role in abstract-state value encoding in vmPFC.

Another recent account proposed that the OFC would be responsible for *credit assignment* i.e. attribute outcomes to specific causes (Walton et al., 2010 [59], 2011 [60]). More precisely, the OFC would be specifically involved for linking feedback value to a particular stimulus in a stream of choices performed over time. The authors examined macaque monkeys' choices before and after focal OFC lesions (lateral OFC). The animals had to decide between three options, ruling out the possibility that their choices would simply reflect perseveration or lack of flexibility (Walton et al., 2010 [59]). The authors argue that the OFC is especially key to guide contingent learning. It does so by selectively attributing an outcome to a particular chosen option alone, and not to options that have been selected simultaneously or close in time.

Despite a growing number of experimental studies implicating OFC in a large variety of computations (value, prediction errors, and their assignment to distinct causes, stimulus/outcome associations encoding ...), the exact role of OFC remains controversial (Stalnaker et al., 2015 [13]; Rudebeck and Murray, 2014 [61]). One of the most influential accounts to date for explaining OFC function across various datasets views OFC a "cognitive map of task space" (Wilson et al., 2014 [62]). OFC would encode the definition of a map of the current task-sets space, allowing for unlearning of old rules to set up new ones, and for guiding behavior in the case of fictive learning (i.e. imagining outcomes that have never been encountered before, simulating possible outcomes, and so on). It relies on the definition and position within a state space. Therefore, simple learning is still possible without OFC but as soon as the task requires more abstract inference, OFC remains necessary. Thus, OFC would acquire and maintain associative representations to guide behavior, in relation with hippocampus and striatum, to which it is connected.

The cingulate cortex is also implicated in value representation for value-based decision-making. In a reward learning task in monkeys, Matsumoto and colleagues found motor/reward contingencies represented in certain medial PFC neurons, separately of visual/motor or visual/reward contingencies representations (Matsumoto et al., 2003 [130]). They recorded cells in the dorsal bank of the ACC that fire to the delivery of juice reward. Based on rewards affective value, monkeys were able to select the most rewarding

stimulus, with neurons in medial PFC thought to underlie this goal-directed behavior. Similarly, Amiez and colleagues found that ACC responses were correlated with the expected quantity of juice, after manipulating the juice probability and amount (Amiez et al., 2006 [131]). The cingulate cortex, especially its anterior part, also responds to the valuation of social information, which is probably rewarding itself (study in monkeys with ACC lesion: Rudebeck et al., 2006 [132]; study in humans with fMRI: Behrens et al., 2008 [133]).

More broadly, ACC neurons appear to encode post-decision variables (Cai and Padoa-Schioppa, 2012 [134]), whereas OFC neurons appear to encode both pre-decision and post-decision variables (Padoa-Schioppa and Assad, 2006 [48]). Another discrepancy between ACC and OFC neurons regards whether value comparison occurs at the level of stimuli (goods) or at the level of actions (motor). In a simple reinforcement task with choice between stimuli or between motor actions, Camille and colleagues examined the behavior of humans subjects with focal lesions centered on dACC or OFC (although for certain subjects damage extended up to preSMA and SMA (Camille et al., 2011 [135]). They revealed that OFC neurons damage implied an inability to sustain the correct choice of stimuli but not of actions (following positive feedback reception). By contrast, damage in dACC led to an inability to sustain the correct choice of actions but not of stimuli (still following positive feedback reception). Similarly, in rats and non-human primates, learning based on stimuli vs. on values can also be distinguished (Ostlund and Balleine, 2007 [136]; Rudebeck et al., 2008 [137]).

Therefore, vmPFC, medial and lateral OFC as well as ACC and dACC form a network implicated in general valuation. Through their connections with premotor and motor systems, these prefrontal regions allow the value comparison process to be converted into an action.

2.2.2.1 Pain and punishments neural correlates

Are negative affective signals such as pain or monetary losses encoded in the same brain regions as positive rewards? Principally, brain responses to negative outcomes are found in insula, MCC and ACC, as well as in thalamus and second somatosensory cortex (Peyron et al., 2000 [138]). More precisely, affective negative value generated by physical pain triggers cingulate activation, generating subsequent cognitive, emotional and motor responses. Using monetary gains and losses, Palminteri and colleagues demonstrate a role for anterior insula in representing the negative affective value of stimuli. Studying learning in patients with brain tumors and patients with Huntington disease, they also

highlight a role for the dorsal striatum in learning to avoid punishments (Palminteri et al., 2012 [139]).

Affective value signals are thus not restricted to ventral regions such as vmPFC.

2.2.3 Conclusion

The combination of behavioral approaches, theoretical models from machine learning and engineering with electrophysiology and imaging studies allowed to understand how subjects learn from rewards, through the affective value that rewards convey. We have seen that a brain network comprising ventral striatum, vmPFC and adjacent medial OFC, PCC and insula supports the valuation of stimuli, items or actions. These affective values representations drive subsequent choices (e.g. dorsal striatum).

Rewards also convey other types of value signals. We have examined the neural substrates of affective values processing and reinforcement learning. However, rewards may convey other types of value signals (O'Doherty, 2014 [54]), such as reward identity, reward sensory features, reward saliency, etc., that is, other signals that are not pure value. In the next chapter, we will focus on the informational value conveyed by rewards, allowing subjects to perform inferences. These inferences about states are at the core of reasoning and decision-making subserved by prefrontal cortex.

Chapter 3

Inferences in human decision-making

Action outcomes convey informational values to adapt behavior in relation to internal mental states. Using informational values, humans are able to perform inferences about states, giving them reasoning capacities and their ability to flexibly adjust their behavior, according to both external contingencies and internal mental states. In this chapter, we will focus on inferential processes that are based on informational values, from psychological, theoretical and cerebral points of view.

3.1 Inferential processes: psychological and theoretical aspects

While humans seem quite rational in their daily life experiences, they often depart from optimality in empirical tests probing for rationality in the sense of formal logic (e.g. Wason's selection task). To explain this discrepancy, tools from the machine learning and artificial intelligence research fields have been imported into cognitive and computational neuroscience in the past few years. The general approach is to "reverse-engineer" the mind, viewing reasoning and learning problems as "computational problems and [viewing] the human mind as a natural computer evolved for solving them" (Tenenbaum et al., 2011 [140]). Here, the notion of inferential processes is not understood in terms of formal logic, but is viewed as probabilistic solutions to a number of concrete problems that humans face. In particular, prior knowledge is usually accompanied by some inherent uncertainty. Priors are not an absolute truth from which one can reason, as compared to premises in formal logic. The theory of Bayesian rationality was developed to re-interpret the apparent irrationality of human choices as compared to formal

logic (Oaksford and Chater, 2009 [141]). In essence, Oaksford and Chater argue that a number of cognitive problems that the brain faces are too complex and therefore computationally intractable. Complex problems put a too high demand on cognitive resources, in terms of memory or processing capacities. But, humans are able to learn very complex models, that they could not possibly be pre-wired for (Pouget et al., 2013 [142]). Rather, Oaksford and Chater suggest that the brain has evolved with cheaper and more efficient solutions to face complex problems: heuristics. A heuristic is a method which does not guarantee to be optimal but is good enough for immediate goals. Therefore, the authors have proposed that the human cognitive system build probabilistic models that are approximate but sufficient [141]. Humans thus seem to use a qualitative probabilistic reasoning, enabling them to deal with real-world uncertain and complex problems. More formally, three structural levels of uncertainty about probabilities can be distinguished. The terminology used to refer to them varies (Yu and Dayan, 2005 [143]; Payzan-LeNestour and Bossaerts, 2011 [144]).

- "Risk", or "noise", or "expected uncertainty". Even if all the probabilities are known, there is a residual hazard. In other words, even after learning all task parameters, in tasks in which reward is probabilistically delivered, there is a remaining uncertainty at each trial about whether a reward will actually be received or not.

- "Ambiguity". This uncertainty level refers to the fact that probabilities can evolve through time. Reward probabilities need to be learnt and estimated through actively sampling the environment. Humans are particularly averse to this type of uncertainty [144].

- "Jumps", or "ignorance". This uncertainty level refers to an abrupt and unpredictable change in external contingencies, such as reversals in learning paradigms. Following the change, the contingencies can reverse to a previously encountered environment, or switch to new contingencies that were never encountered before. This uncertainty level relates to situations in which the agent does not even know the space of all possible states.

The three levels correspond to (1) uncertain, (2) changing and (3) open-ended environments respectively. Expected and unexpected uncertainty have been used to refer to either levels (1) and (2) [143] or levels (2) and (3) [144]. In experiments, humans are not able to accurately discriminate between these structural levels of uncertainty (Payzan-LeNestour and Bossaerts, 2011 [144]).

3.1.1 Probabilistic models of learning and reasoning

The brain relies on inductive systems for learning. An inductive system is a form of reasoning which proposes general laws on the basis of few particular observations, in a probabilistic way. It means that the generalization is not necessary true, unlike in deductive reasoning, but is assumed to be true after a number of repeated observations. The generalization go beyond the available observations, but remains accompanied with some uncertainty. This inductive capacity is thought to be based on the human tendency to look for regularities (Yu and Cohen, 2009 [145]). Inductive learning allows both children and adults to generalize knowledge on the basis of very few observations (Tenenbaum et al., 2011 [140]; Collins and Frank, 2013 [37]). For example, learning an abstract concept or a word meaning requires to generalize from sparse and uncertain information (Griffiths et al., 2010 [146]). In this review, Griffiths and colleagues defend the top-down approach of probabilistic models to understand cognition, as opposed to a bottom-up connectionist approach, which studies neural networks and looks at the emerging properties they present. The top-down approach enables that qualitatively different types of representations can be used for learning in different domains. Moreover, the probabilistic models top-down approach allows to integrate pieces of information such as verbal instructions, and to swiftly adapt learning consequently. In contrast, a bottom-up connectionist model would have difficulty rearranging rapidly for including information such as a verbal instruction. Therefore, probabilistic models of cognition describe human inductive learning and reasoning through Bayesian inference [140].

3.1.2 Bayesian inference

The core of probabilistic reasoning, involved for instance in solving inductive problems, is expressed in Bayes rule:

$$p(h|d) \propto p(d|h)p(h) \tag{3.1}$$

with $p(h|d)$ corresponds to the posterior, i.e. the probability (**belief**) that the hypothesis (h) is true given the data (d). The posterior is actually a probability distribution over all hypotheses, after observing the data. It represents how likely is each hypothesis after observing the data. Next, $p(h)$ corresponds to the prior distribution over all hypotheses before observing the data. It captures the degree to which the agent is biased towards one hypothesis or the other beforehand, independently of observing the data (i.e. inductive biases). Finally, the likelihood $p(d|h)$ is the probability of observing the data knowing that the hypothesis h is true. It represents how well the hypothesis h fits the data. Note: here the constant term $p(d)$, global probability of observing the data is not shown. It permits that the posterior probabilities sum to 1 (Figure 3.1).

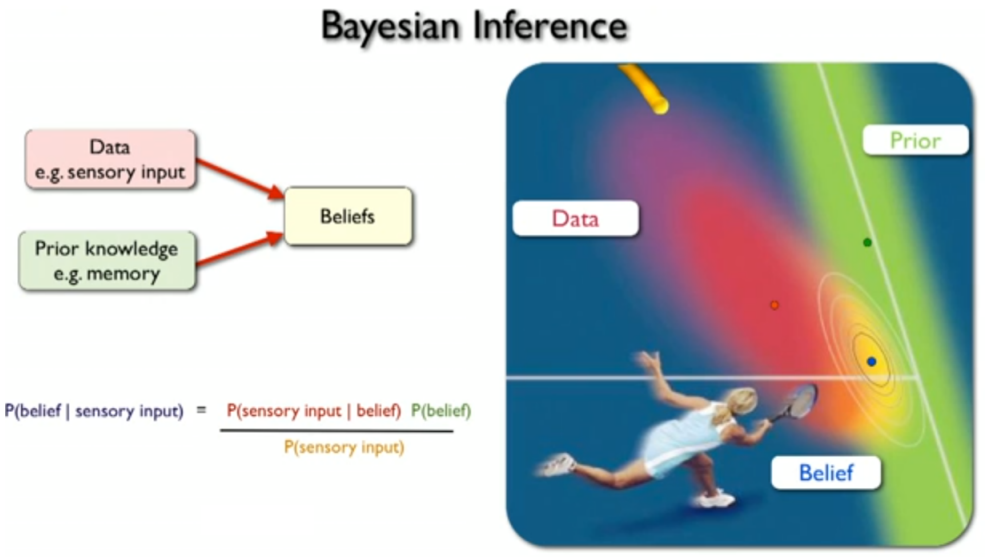


FIGURE 3.1: Illustrative example of Bayesian inference to compute a posterior belief predicting where the tennis ball is going to fall, combining prior expectations with sensory evidence (likelihood) (reproduced from Wolpert, 2013).

Thus, the Bayesian framework explicits the way in which evidence can be incorporated into prior knowledge in causal learning. The prior and posterior probability distributions represent the **degree of belief** in a statement. Importantly, the Bayesian approach sees probabilities as a scalar interpretation of a degree of knowledge, of confidence in favor of a hypothesis. The Bayesian approach is fruitful when data are rare or sparse, and is less sensitive to data volatility. A number of learning problems can be explained as Bayesian inferences, such as language acquisition (Goldwater et al., 2009 [147]), property induction (Kemp et al., 2009 [148]), causal attribution from observations, etc. It allows generalization. For example regarding language acquisition, children ability to extrapolate word meaning and language structure largely outstrips the available evidence they have (Xu and Tenenbaum, 2007 [149]). Regarding causal learning, observations of events co-occurrence are constrained by abstract prior knowledge. Indeed, the sensory evidence observed (likelihood) is often noisy and uncertain (e.g. ambiguous speech signal or Figure 3.2).

Therefore, prior knowledge appears essential for constraining the decision problem. This abstract knowledge can be learnt from experience and used for subsequently acquiring more specific knowledge (Tenenbaum et al., 2011 [140]).

Another proposal by Friston and colleagues describes inferential problems under the free energy minimization principle (Friston et al., 2013 [150]). The concept of free energy is a measure of statistical probability distributions. It comes from thermodynamics and refers to the difference between a system's energy and entropy (i.e. the amount of energy

FIGURE 3.2: In real-world decisions, the sensory evidence (likelihood) available to update beliefs using Bayesian inference can be noisy and ambiguous.

directly available for producing work). The system can be a biological organism or organ, e.g. the brain, which changes to minimize its free energy (Helmholtz, 1860; Friston et al., 2006 [151]), allowing to react to changes in the environment. Its internal models could have been selected within a population, with heritable priors. More precisely, action selection is viewed as an active inference problem, in which the agent chooses the action that minimizes free energy (Friston et al., 2006 [151]). Simply put, minimizing free energy comes back to minimizing the distance (Kullback-Leibler divergence) between two probability distributions, the exact and approximate posteriors. Action is then sampled from posterior beliefs about control. Therefore, selection is not based on action values, it is free energy minimization that is driving choice, and values arise as a consequence of choice (Friston, 2010 [152]). The states that are more frequently occupied become more valuable. Active inference and free energy minimization have been developed mostly in the context of perceptual decision-making. Exact Bayesian inference cannot be realized for being computationally intractable, but approximate Bayesian inference can be performed, leading to bounded rationality. Beliefs entail a notion of precision, to balance the influence of prior (biases) and sensory evidence (likelihood).

3.1.3 Possible limits of the Bayesian approach.

One of the main limits of the Bayesian approach is that it hypothesizes that the system can have an exhaustive representation of all possible states on which Bayesian inference is subsequently performed. This is expressed in the constant term used for normalization in Bayes rule, which supposes that the states full partition is known. This is rarely the case in real life open-ended environments. Nevertheless, Bayesian inferential learning is efficient when the generative model i.e. the state space structure is known. Notably, there is a difference between inferential processes and model-based reinforcement learning. Although model-based RL includes a notion of state, in model-based RL there is no

Bayesian inference, no learning of the structure. In model-based reinforcement learning, the state space structure within which learning is effected is supposed to be known.

Other lines of evidence suggest that real-world learning problems present too many dimensions so that simple Bayesian inferences cannot be computationally performed. Indeed, stimuli present various sensory dimensions, or sometimes unrelated events coincidentally co-occur, without a link to be learnt. In a recent study, Niv and colleagues showed that a statistically optimal Bayesian model did not explain behavior on a multi-dimensional RL task. More generally, attention has to be directed to specific dimensions of our senses that are specifically relevant for learning (Niv et al., 2015 [153]; Geana and Niv, 2015 [154]). Moreover, since Bayesian inference presents a high computational cost, it is unlikely that the whole human brain entirely operates as a Bayesian system (Eckstein et al., 2004 [155]). Model-free reinforcement learning should be preferred in situations in which the Bayesian approach does not provide much extra value. For efficiency, the brain must trade-off cost and performance (O'Reilly et al., 2012 [156]).

In some cases, inferential processes permit the extrapolation of outcomes to unchosen option(s). This fictive learning infers what would have been the outcome had the choice been different. The phenomenon of learning from unchosen actions, namely, counterfactual learning, for example leads humans to experience regret (Coricelli et al., 2005 [157]; Coricelli and Rustichini, 2009 [158]). Counterfactual information reinforces the dependence on context (frame of reference) for evaluating rewards and punishments (Palminteri et al., 2015 [159]).

3.1.4 Application of Bayesian inference models to learning and decision-making

A recent work in our team proposed an inferential model to explain how humans learn, adjust, create and retrieve behavioral strategies in changing, variable and open-ended environments (Collins and Koechlin, 2012 [160]). Here, "behavioral strategy" is understood as "task-set" i.e. a representation of a mapping between stimulus, action and outcome.

The PROBE model [160] responds to a main inferential problem: in real life open-ended environments, the range of possible behavioral strategies can expand infinitely. In that case, optimal Bayesian inference is described by Dirichlet process mixtures, such that it rapidly becomes intractable. This computational complexity has shaped the inferential processes evolution in prefrontal cortex. The PROBE model constitutes a biologically plausible approximation of Dirichlet process mixtures, which accounts for the human

executive function limits. It combines reinforcement learning, limited Bayesian inference and hypothesis testing to arbitrate between adjusting, switching and creating task-sets.

The PROBE model [160] proposes that human executive function consists of a monitoring system of each task-set reliability i.e. the posterior probability of the task-set currently being the most accurate one to guide decisions. On the basis of the outcomes received using a certain task-set's predictions, the task-set's reliability is inferred. Absolute reliability is then assessed using hypothesis testing. If the task-set is more reliable than unreliable, it is chosen to drive action selection. While it remains more reliable than unreliable, it is maintained to guide choices (exploitation phase). By contrast, if the task-set is more unreliable than reliable, the decision-maker switches to the use of another task-set. This switch marks the beginning of an exploration phase. Entering an exploration phase, the decision-maker can either retrieve a task-set from the working memory buffer, or temporarily create a new task-set. The new task-set creation is partly under long-term memory influence, which records the frequency of past use of each task-set for action selection. The newly created task-set can then be either confirmed or discarded, according to its success in accurately predicting outcomes. Once a task-set is retrieved or confirmed, going back to an exploitation phase, it is adjusted online through reinforcement learning. Lastly, the working memory buffer capacity appears to be limited to the monitoring of about three task-sets simultaneously. However, inter-individual variations were observed, regarding individual working memory capacity, and regarding subjects' tendency to exploit vs. explore [160].

Therefore, the PROBE model relies on Bayesian inference for *hypothesis-testing*. Indeed, the choice of the task-set driving action selection is based on testing the current task-set absolute reliability. The notion of reliability refers to the degree of belief of being in a particular state of the world. Reliability measures how much the current task-set matches the current external contingencies, in other words, what is the best task-set to exploit now. Formally, reliability corresponds to the degree of **belief about how actions map onto outcome contingencies**. This belief allows to monitor and adapt behavior in relation to internal mental states. Monitoring processes define the dynamic evaluation of a series of mental representations maintained in working memory, that subsequently drive behavior. Monitoring sometimes also refers to the attentional processes towards working memory content, which make possible this online evaluation (Schraw, 1998 [161]).

Conclusion. We have seen that the Bayesian framework provides a convincing theoretical account to explain human learning. Probabilistic models of cognition explain how humans are able to make Bayesian inference and generalize from limited experience. Qualitative probabilistic reasoning enables humans to deal with real-world uncertain

and complex problems. Therefore, humans do not only learn and decide on the basis of observed rewards but use prior knowledge to guide their decisions. Prior knowledge variability across people might explain why medical or law facts can be differentially incorporated, and explain why, although facing the same evidence, different people would make different decisions. In the next section, we will examine the neural mechanisms subserving inferential processes.

3.2 Inferential processes: cerebral aspects

What are the brain regions underpinning Bayesian inference and reasoning abilities in humans?

3.2.1 Model-based neuro-imaging

Behrens and colleagues analytically developed a Bayesian model which consist of the optimal behavior in a probabilistic reversal learning task (Behrens et al., 2007 [162]). This study nicely introduces the notion of environment volatility. In their task, subjects had to make a decision between two options providing stochastic rewards. The most frequently rewarded option reversed from time to time, often (volatile period) or rarely (stable period). The proposed Bayesian model learns online the reward probabilities associated with each option, and infers in parallel the rate of reward probability changes, i.e. volatility. The volatility itself is controlled by an additional parameter. At each trial, action is stochastically selected according to the largest of the two option values (reward probability x reward magnitude). According to the Bayesian model, the agent is learning online all task variables (Figure 3.3).

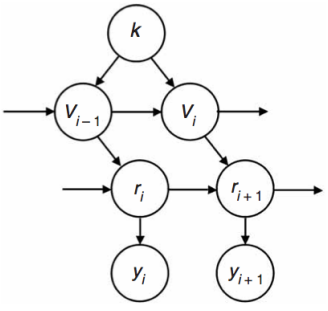


FIGURE 3.3: Graphical description of the Bayesian model including volatility-related modulation (reproduced from Behrens et al., 2007).

Neural correlates of the inferential model were found in vmPFC and PCC for action values, whereas the volatility was found to be encoded in ACC, dorsally, specifically

during the time period following feedback reception [162]. Critically, subjects were able to modulate their learning rate given the environment volatility. In volatile periods, of high uncertainty, subjects gave more importance to the recent past outcomes. However, in more stable periods, subjects took into account a larger reward history. Therefore, Bayesian inference enables to modulate the weight given to each new piece of information.

Similarly, Donoso and colleagues analyzed the PROBE model (Collins and Koechlin, 2012 [160]) neural implementation using a model-based fMRI approach (Donoso et al., 2014 [94]). The model architecture is further described in the previous section. They revealed neural correlates of the actor task-set reliability in vmPFC and perigenual ACC, whereas the reliabilities of alternative task-sets (best and second-best alternatives) were found in bilateral frontopolar cortex (Figure 3.4).

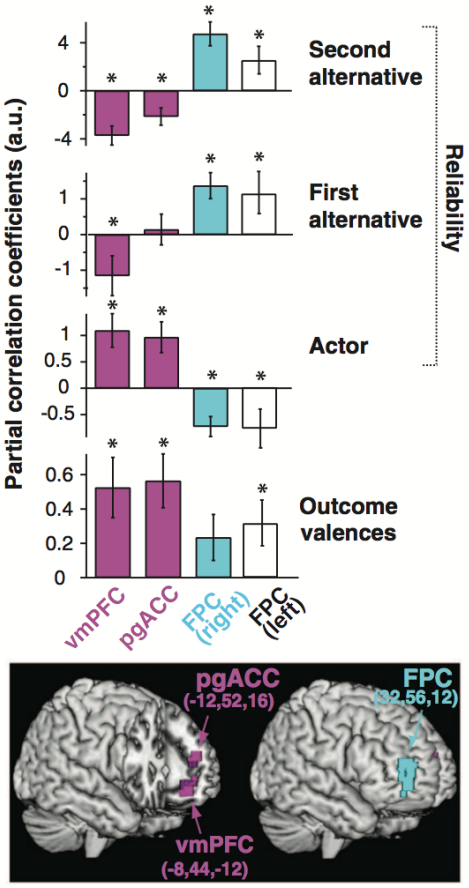


FIGURE 3.4: Neural correlates of reliability signals according to the PROBE model (reproduced from Donoso et al., 2014).

In addition, dACC identifies when the actor task-set turns unreliable, prompting an exploration period, with the retrieval of a previous task-set or the creation of a probe task-set that will be later confirmed or discarded. Moreover, the dACC implication in provoking exploration has been shown before, in the context of foraging decisions.

Neuronal recordings in macaque monkeys showed that dACC activity increased up to a certain threshold, that triggered the decision to leave a food patch to explore other resources, in a mechanism that resembles drift-diffusion models (Hayden et al., 2011 [81]).

In addition, ventral striatum was identified to respond to confirmation events, meaning the validation of a recently created task-set to guide decisions (Donoso et al., 2014 [94]). Importantly, the region was identified thanks to specific algorithmic events predicted by the PROBE model (confirmation events), that could not have been located in time otherwise. This is a piece of evidence showing that vmPFC activity was consistent with an inferential model.

However, we have seen in the previous chapter that vmPFC is also responsible for encoding the affective value of stimuli driving choices. Therefore, is vmPFC activity more consistent with an abstract-state-based model than with an affective value-based model?

3.2.2 A role for vmPFC in inference

A convincing piece of evidence came from an elegant study by Hampton and colleagues (Hampton et al., 2006 [56]). Using model-based fMRI, the authors tested whether vmPFC activity was better explained by a state-based inferential model or by a reinforcement learning model. In their probabilistic reversal learning task, subjects had to choose between two stimuli, for which the reward contingencies were anti-correlated. The task was quite hard because soon after subjects identified the "good" option to choose, the two options reversed. The intuitive prediction corresponds to the following reasoning. After a negative outcome reception or a series of negative outcomes, subjects more often switched to the other option. When subjects decide to switch, reinforcement learning and state-based model make different predictions. If subjects use a reinforcement learning model, the value of the newly chosen option should be low, because it was low the last time the subject chose it and subsequently abandoned it. By contrast, if subjects use a state-based model, the value of the newly chosen option should be high, because the subject inferred the underlying task structure. She has inferred that the two options are anti-correlated (if one is low, the other is high), so when she abandoned a low-valued option, she knows that the value of the newly chosen option should be high. The authors observed that qualitatively, BOLD activity in vmPFC was rather consistent with a state-based model (Figure 9.1). We will see that our results are in line with these data. However, their protocol limit lies in the use of binary rewards (win/lose); they did not parametrically modulate rewards (no notion of affective value).

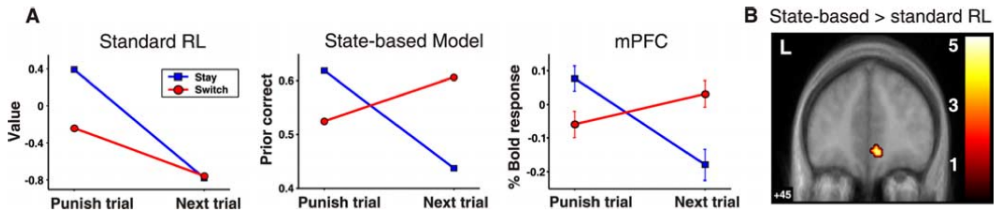


FIGURE 3.5: In switch trials (red lines), fMRI activity in vmPFC was rather consistent with a state-based model (belief values) than with a reinforcement learning model (affective values) (reproduced from Hampton et al., 2006).

3.2.3 The medial PFC functional architecture in decision-making

Taken together, these data challenge the original view of medial PFC as the more ventral the more affective, the more dorsal the more "cognitive" (Bush et al., 2000 [163]). Indeed, we have just seen that the ventral part of medial PFC (vmPFC) activity was consistent with inferential models, based on abstract state-based inferences or reliability signals computed through Bayesian inference. However, we have also seen in the previous chapter that the vmPFC activity encoded subjective affective values, as supported by a large number of empirical studies in animal and humans. Similarly, we have discussed the dACC role in Bayesian inference. For example, dACC is involved in inferring when to switch to an alternative course of action, or exploring. However, we have seen in the previous chapter that dACC are also involved in the processing of affective primary value signals such as pain, and general emotional experience. Moreover, other authors have challenged the original view of "ventral affective" vs. "dorsal cognitive", gathering evidence that both vmPFC and dACC are involved in regulating affective responses (Etkin et al., 2011 [164]; Shackman et al., 2011 [165]).

The medial PFC functional organization regarding decision-making thus remains unclear. Rather, the experimental evidence available up to now suggests a ventral/dorsal functional architecture in medial PFC regarding "stay" decisions vs. "switch" decisions. Broadly, vmPFC would be engaged when a decision to stay is made: exploiting the same option, monitoring the current state, in the default mode network, etc. In contrast, dACC, would be recruited when a decision to switch is made: leave a default option, switch task-set, explore, etc. (Boorman et al., 2013 [87]). vmPFC would encode the tendency to repeat choices while dACC would be engaged when leaving a default behavior.

Conclusion. In this section, we scrutinized the cerebral bases of Bayesian inference in human decision-making. Several brain networks support the implementation of probabilistic models of cognition, which allows humans to cope with uncertain and complex

learning and decision-making problems. More generally, dACC seems to be engaged when a decision to switch is about to be made: leave a default option, switch choices, explore... Brain inferential systems thus would allow to rapidly detect a reversal or a change in environmental contingencies and implement the necessary adaptations, in a rapid and flexible manner. By contrast, choice models based on affective values present a continuous but slower adaptation, and cannot flexibly switch to a new behavior as soon as a change in external contingencies is identified (Keramati et al., 2011 [166]). The ability to flexibly react and adapt action in relation to internal mental states is at the core of human prefrontal executive function.

Chapter 4

Research question

We have seen that executive control relates to the human ability to monitor and flexibly adapt behavior in relation to internal mental states, crucially implicating prefrontal cortex. Specifically, executive control and decision-making rely on evaluating action outcomes to adjust immediate and future action (Chapter 1).

Actions can be reinforced or devaluated according to the outcomes affective value, through conditioning, as formalized in reinforcement learning theories. A large body of neural data from rats, monkeys and humans involves notably basal ganglia and medial prefrontal cortex in affective values processing (Chapter 2).

In addition, action outcomes convey information to adapt behavior in relation to internal beliefs, relying on Bayesian inference. Inferential mechanisms are subserved by prefrontal cortex and allow learning and generalizing from outcomes through belief updating (Chapter 3).

Accordingly, we have been working on the idea that action outcomes convey two major types of value signals:

- **Affective values**, representing the action outcomes valuation according to subjective preferences, and stemming from reinforcement learning. Here, *affective value* is understood as reward magnitude. Unlike its common meaning, the term *affective* here refers to the motivational properties of outcomes for action, rather than emotional properties.
- **Belief values**, about how actions map onto outcome contingencies, and relating to Bayesian inference.

To our knowledge, previous experimental paradigms have confounded these two types of value signals. Indeed, in natural settings, obtaining rewards of high affective value

usually *informs* about more appropriate choices. In other words, receiving a rewarding outcome naturally increases the belief that the chosen action was the most appropriate one. However, how these two signals contribute to decision-making remains unclear. In this PhD work, we investigate whether this dissociation is behaviorally meaningful, and whether the two signals, beliefs and affective values, have distinct neural bases.

To address this question, we developed a series of behavioral experiments in tandem with computational modeling and functional magnetic resonance imaging in healthy human subjects. More precisely, the key feature of the probabilistic reversal-learning tasks presented here was to decorrelate affective values from belief values using stochastic and changing reward structures. We built a computational model that establishes the functional and computational foundations of such dissociation. The model combines two parallel systems: reinforcement learning, dealing with affective values, and Bayesian inference, dealing with belief values. The model better accounted for subjects' behavior than many other alternative models. Critically, neural data revealed a double-dissociation between ventromedial prefrontal cortex and midcingulate cortex (MCC) regarding choice-independent effects, with ventromedial prefrontal cortex being specific of beliefs while midcingulate cortex was specific of affective values.

Concretely, we present in Chapters 5 and 6 three data sets corresponding to three variants of a behavioral study (Protocol A). In Chapters 7 and 8, we present two pilot studies and an fMRI study (Protocol B). Chapter 9 concludes with a general discussion of the results, in light of the recent literature regarding prefrontal cortex and decision-making.

Chapter 5

Protocol A: Decorrelate affective value from information of outcomes

We present a series of three probabilistic reversal learning tasks, involving stochastic and changing reward structures, aiming at dissociating affective values from belief values of action outcomes.

5.1 Experiment 1

In a first experiment, we manipulated stochastic and changing reward distributions to de-correlate reward affective value from belief value.

5.1.1 Experimental design

Subjects had to make a decision between two stimuli (Figure 5.1). After choice, they received an outcome among the possible values 1, 2, 5, 8 and 9 Euros. Stimuli were simple shapes (circle or square) or letters (A or B).

Each of them represented a one-armed bandit. A one-armed bandit corresponds to a slot machine which delivers rewards probabilistically. Crucially, one of the two bandits triggered on average a higher amount of reward (6.23 Euros vs. 3.77 Euros per trial). The two-armed bandit were anti-correlated (Figure 5.2). Without any sensory cues, the mapping between stimuli and bandit shifted after an unpredictable number of trials. This transition from an episode to the next one was called a reversal.

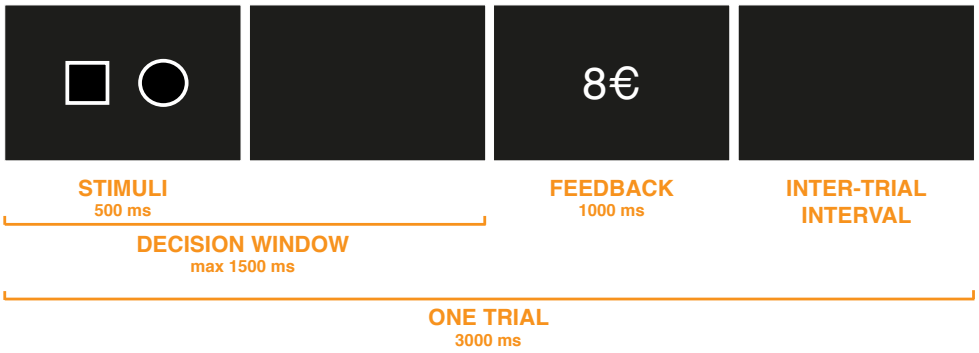


FIGURE 5.1: Probabilistic reversal learning task: trial structure and timing.

The key feature of the experiment was the reward distributions underlying each bandit. Critically, we designed, by perturbing an exponential distribution, a bimodal reward distribution (Figure 5.2) in which the affective value of reward was no longer correlated with the *correctness* (i.e. choosing the highest rewarded bandit on average).

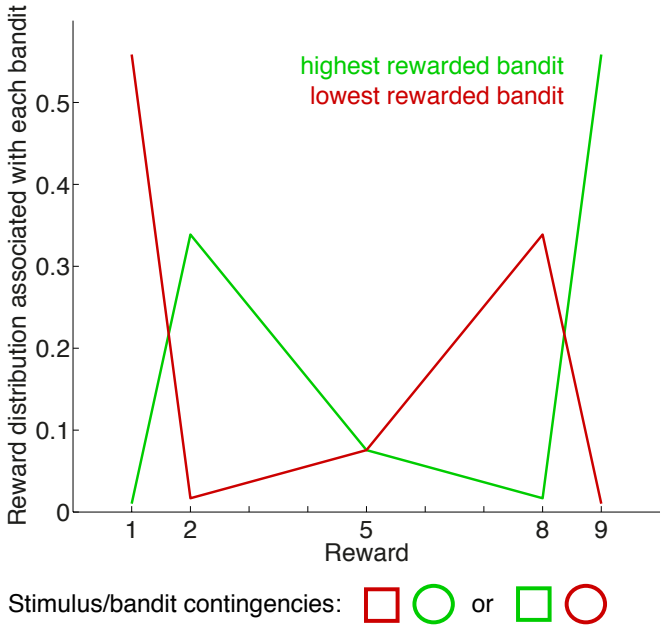


FIGURE 5.2: Reward distributions underlying each bandit.

The intuition behind the "trick" of the experiment was the following. Receiving 2 Euros had a low affective value but informed about having chosen the highest rewarded bandit (= better choice on the long run). By contrast, receiving 8 Euros had a high affective value but informed about having chosen the lowest rewarded bandit (= worse choice on the long run). Receiving 5 Euros was uninformative. Finally, receiving 9 Euros (respectively 1 Euro) was highly informative about having chosen the highest (respectively

lowest) rewarded bandit i.e. consistency between affective and informational values in the case of 9 and 1 Euros outcomes. The distribution and value scale were chosen to maximize the differences between the predictions of a RL (Rescorla-Wagner rule) model and a Bayesian model. The models will be described in the next section.

Subjects were asked to learn which of the two stimuli was the highest rewarded bandit, knowing that the best of the two stimuli could change over time, and to respond in order to win as much money as possible.

Each subject completed two sessions, which differed only in the first 5 minutes. In one session, participants were primed to use reinforcement learning, whereas in the other session, participants were primed to use Bayesian inference. We designed these "primes" assuming that they would modulate subjects' strategies.

For the "RL prime", the reward distribution was an exponential-like distribution (Figure 5.3). On this kind of distribution, ideal reinforcement and ideal Bayesian learner would perform equally well. Given the cognitive cost of setting up a Bayesian strategy, which is more sophisticated, we expected subjects to use a simple and more parsimonious reinforcement learning strategy on this prime.

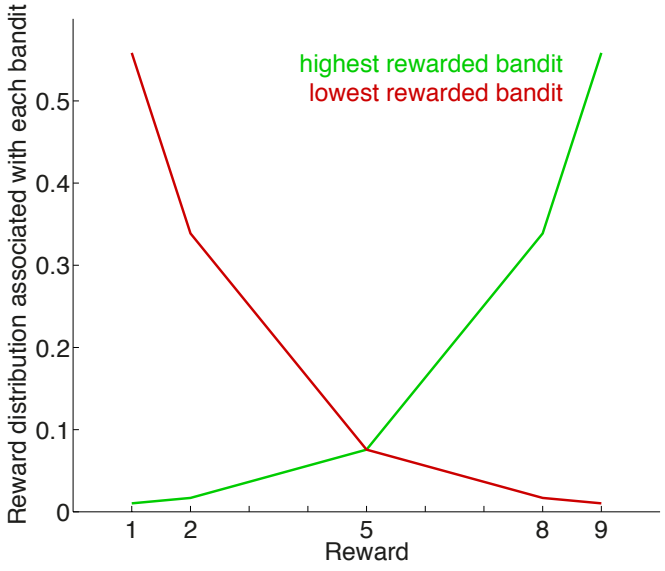


FIGURE 5.3: Reward distributions underlying each bandit during "RL prime".

For the "Bayesian prime", reward distributions for both bandits were uniform i.e. random distribution of all possible outcome values (Figure 5.4). In other words, there was no good or bad choice and both bandits were strictly equivalent. We hypothesized that there will not be motivational problems since it was only on the first five minutes. Because no structure could be inferred on this distribution, and because reinforcement

learning would be inefficient, we expected subjects to rather try a Bayesian strategy on such a distribution (e.g. try to find a pattern or a structure in the task).

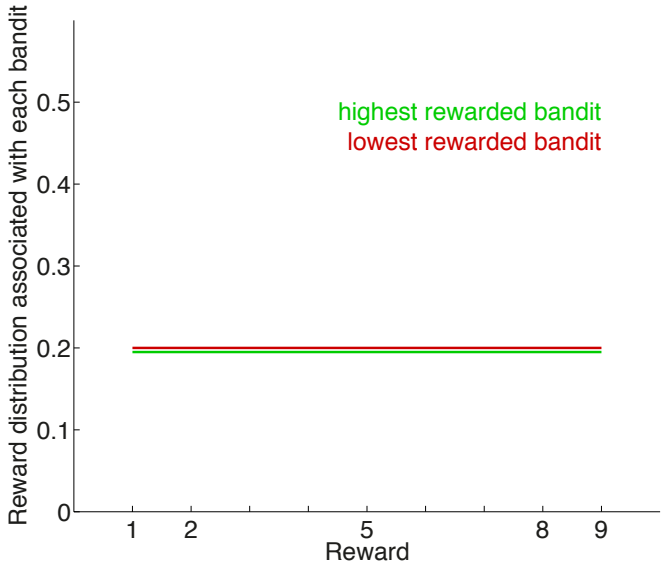


FIGURE 5.4: Reward distributions underlying each bandit during "Bayesian prime".

Because of this "priming", there was no training on the bimodal distribution. In the end, this "priming" manipulation towards a particular strategy did not significantly modulate subjects behavior on the subsequent bimodal distribution. Therefore I will focus in the rest of the chapter on analyzing subjects' behavior on the bimodal distribution.

At the end of both sessions, an informal debriefing was done with each participants. After comments from their own initiative, subjects were asked, among other questions, whether they used a particular strategy or rule to respond, and whether they noticed any regularities or organization within the experiment. The original debriefing is provided in Appendix A.

5.1.2 Randomization

The task was fully counterbalanced and pseudo-randomized. Each stimulus appeared an equal number of times in a random order. Each stimulus was associated to a motor action (key on computer keyboard) as many times as to the other action. The highest rewarded bandit was associated to a stimulus as many times as to the other stimulus. Each session included 40 episodes (= 39 reversals), an episode being a cluster of trials using the same stimulus/bandit association. Episodes could have various lengths (either 16, 22, 26, 30 or 36 trials. The frequency of each episode length apparition followed a Gaussian centered at 26 trials. Overall, each session included 1040 trials, 104 trials on

prime and 936 trials on bimodal distribution. The order of sessions and the stimuli set (circle/square or A/B) used for each session were counterbalanced across subjects.

5.1.3 Experiment presentation

All stimuli were presented using PsychToolBox [167] and appeared on a uniform black background as shown in Figure 5.1. Each session was divided by four breaks. The five parts thus formed were roughly equally long, but subjects were told that there was no particular meaning to the moment of the break apparition, such as they would not infer any rule or pattern related to the breaks. Each trial lasted 3000 ms. Stimuli were displayed during 500 ms but subjects could respond within 1500 ms after stimuli apparition. If they did not respond within 1500 ms, the trial was lost. They had to respond by pressing one of two computer keys. 100 ms after they respond, feedback was given: the value of the obtained reward was displayed in the center of the screen during 1000 ms. Finally, a 400 ms inter-trial interval separated each of the trials.

5.1.4 Participants

25 healthy individuals (12 males, aged 18-25 years) with normal or corrected-to-normal vision, no general medical, neurological, psychiatric or addictive history were recruited in Paris, France through an internet database (http://expesciences.risc.cnrs.fr/). They all gave written informed consent (approved by the French National Ethics Committee) during a medical interview with our on-site physician for their participation in two behavioral sessions which took place on two separate days (yielding 1040 trials over approximately 52 minutes of testing per session). Subjects were paid 40 Euros. They received written instructions about the task and were instructed that payoffs could vary according to their own performance, to hopefully maintain a high enough motivation along the whole session.

5.1.5 Statistical analysis

Behavioral analyses were performed under MATLAB R2011a. Based on a performance at chance level during the "RL prime", we excluded two subjects. Data from all remaining subjects (N = 23) were pooled for behavioral analyses. We checked whether no effects of prime, of order of sessions, of age, of sex, of years of education were observed. Choice proportion of the highest rewarded bandit was calculated for each participant over the course of an episode and averaged across subjects (Figure 6.1). Moreover, we computed the choice proportion of the highest rewarded bandit *per episode* to investigate

whether there was a progression in performance over the whole experiment (Figure 6.2). Also, we compared the performance between first half and second half of experiment, assuming that subjects might have better established a response strategy on the second half of experiment.

Theoretical simulations (not shown) revealed that, on the best tuning of this particular bimodal reward distribution, an ideal Bayesian learner (i.e. with optimal parameters and no decision noise on action selection) would get on average more reward than a reinforcement learner. According to the following intuitive reasoning, we hypothesized that RL and Bayesian learning will be distinguishable on this task. Consider a subject who picked the lowest rewarded bandit on a given trial, for which she obtained 8 Euros. (1) If she performed according to RL on the task, integrating only reward affective value (i.e. reward magnitude), she would wrongly assume that she chose the highest rewarded bandit, since reward magnitude was high. Therefore, she should choose the same bandit again on the next trial. (2) If she performed according to Bayesian inference on the task, she was able to infer the reward distribution behind each bandit and use it to respond. So, she understood that she picked the lowest rewarded bandit, even though she locally got a high-magnitude reward. Therefore, she would switch bandit on the next trial. Thus, we hypothesized that computing the switch/stay behavior after a reward of 8 Euros (and, symmetrically, of 2 Euros) would allow us to discriminate between subjects who did or did not infer the underlying task structure. Essentially, we computed the stay trials proportion after each obtained reward, averaged over all subjects (Figure 6.3). Furthermore, we reproduced this stay/switch analysis comparing between first and second half of sessions, to investigate possible meta-learning of the task structure over the course of a session. Also, we compared stay/switch behavior between the onset of an episode (just after a reversal) and end of episode, under the hypothesis that after a reversal, subjects might use simpler strategies (RL) whereas at the end of an episode, i.e. in a more stable period, they might use more sophisticated strategies (Bayesian).

However, the above intuitive reasoning about the task was based on only the previous outcome. Indeed, the learning rate parameter fits in RL models (see below) confirmed that subjects based their choices not only on the previous outcome but taking into account a larger reward history.

5.1.6 Computational modeling

To establish the functional and computational foundations of the dissociation between affective and informational values, we examined and developed mathematical models of

learning and decision. The aim of such cognitive modeling is to understand the mechanistic computations underlying behavior, i.e. hidden variables that are not directly visible in behavioral data. We emphasize the importance of testing various realistic alternative models, that are challenging candidate models. A model accuracy in explaining subjects' behavior is always relative. The selected model is the best only among a limited number of candidate models, which are never exhaustive.

5.1.6.1 Reinforcement learning model

We studied each participant's trial-to-trial choices by a reinforcement learning model [109]. This model hypothesizes that expected values Q_t for each bandit were learnt from observations of rewards using the following equations.

Standard RL model. In the standard RL version, only the Q value of the chosen action was updated, according to:

$$Q_{t+1} = \begin{cases} Q_t + \alpha(r_t - Q_t) & \text{receiving a reward } r_t \text{ for chosen action} \\ Q_t & \text{for unchosen action,} \end{cases} \tag{5.1}$$

in which α was a learning rate parameter. At the beginning of each session, Q values were initialized at their mean value (5 Euros), given subjects had no reason to prefer any of the two stimuli. Fitting the initial Q values as a free parameter did not significantly improve the model. Importantly, the RL model consisted in a continuous trial-by-trial update; it assumed no task structure, specifically, no structure related to the reversals.

Normalized RL model. In the normalized RL version, both chosen and unchosen Q values were updated at each time step, according to:

$$Q_{t+1} = \begin{cases} Q_t + \alpha(r_t - Q_t) & \text{receiving a reward } r_t \text{ for chosen action} \\ Q_t + \alpha(10 - r_t - Q_t) & \text{for unchosen action,} \end{cases} \tag{5.2}$$

More precisely, the normalized RL assumed that if a reward r_t was received, choosing the other option would have led to a reward of 10 Euros$-r_t$, knowing that the reward scale was centered on 5 Euros; Implicitly, this model assumed an underlying structure in the task, which was that the two bandits would be opposite. But it assumed no structure related to reversals.

5.1.6.2 Bayesian inference model

The underlying generative model of the task (Figure 5.5) corresponded to the statistically optimal model. It consisted of a hidden Markov model i.e. all variables of the current

state depended only on the previous state; there was no backward inference. The task of the decision-maker was to figure out the hidden state.

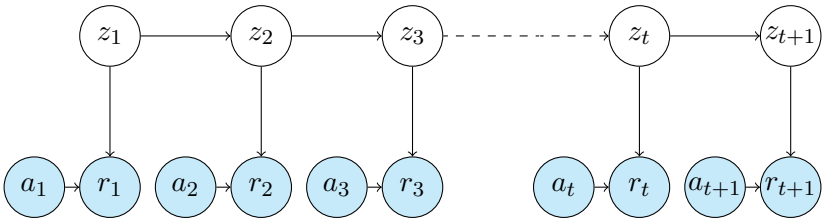


FIGURE 5.5: Experiment 1, 2 and 3: Generative model of the task. z_t: underlying hidden state; a_t: action performed; r_t: reward received.

According to this Bayesian model, the agent maintained a **belief** about how stimuli mapped onto reward distributions (stimulus/bandit mapping). Choice was stochastically selected according to the belief $p(z_{t+1} = 1|a_t, r_t)$ (see paragraph Action selection). After choice, a feedback was observed and led to two inference steps.

(1) The belief was updated with given feedback received r_t and given reward distributions $p(r_t|z_t = 1)$ and $p(r_t|z_t = 0)$ presented in Figure 5.2 according to Bayes rule:

$$\begin{aligned} p(z_t = 1|a_{1:t-1}, r_{1:t-1}) &\propto p(r_t|z_t = 1) \times p(z_t = 1|a_{1:t-1}, r_{1:t-1}) \\ p(z_t = 0|a_{1:t-1}, r_{1:t-1}) &\propto p(r_t|z_t = 0) \times p(z_t = 0|a_{1:t-1}, r_{1:t-1}) \end{aligned} \tag{5.3}$$

With both expressions at hand, these probabilities were normalized to achieve $p(z_t = 1|\dots) + p(z_t = 0|\dots) = 1$.

(2) The volatility ν corresponded to the probability that the hidden state had changed: $p(z_{t+1}|z_t)$. A transition step was performed using volatility:

$$p(z_{t+1} = 1|a_t, r_t) = (1 - \nu) \times p(z_t = 1|a_t, r_t) + \nu \times p(z_t = 0|a_t, r_t) \tag{5.4}$$

An analogous expression holds for $p(z_{z+1} = 0|\dots)$.

Furthermore, the volatility ν was constant across trials, meaning that the volatility did not vary across trials, whether close or far from reversals.

At the beginning of each session, beliefs were initialized at their mean value (0.5, unbiased ideal observer), given subjects had no reason to prefer any of the two stimuli. Fitting the initial belief as a free parameter did not significantly improve the model. The weakness of this Bayesian model was that it assumed that the decision-maker had knowledge of the reward distributions (unlike the Bayesian model presented in the next

section that learnt trial-by-trial the reward distributions). In practice, there was a training beforehand, and fits were reproduced on the second half of trials to ensure subjects had enough time to sample and learn the reward distributions.

Importantly, the Bayesian model used knowledge of the task structure i.e. the reward distributions underlying each bandit, in order to adapt faster when a reversal occurred.

5.1.6.3 Bayesian inference model with online learning

The above Bayesian model had knowledge of the reward distributions beforehand and use them to learn the stimulus/bandit mapping. By contrast, this Bayesian model learnt online the reward distributions, meaning that, it had to learn in parallel (1) the stimulus/bandit mapping and (2) the reward distributions. To simplify the learning problem, we approximate the model by including only forward inference. The model learned the reward distributions from the space of all possible distributions, with an exponential-like prior on distributions. There was no backward revision on the previous trials history. The free parameters were still β and ϵ for softmax and ν (volatility).

5.1.6.4 Decay model

In addition, we implemented a version of the above Bayesian model with online learning with a temporal decay, to consider possible memory loss during reward distributions learning. There was an additional decay parameter on distributions that exponentially degraded learning over past trials. The decay parameter was allowed to vary across subjects but was assumed to be constant over a session. We tested two versions of this decay: one in which the decay affected multiplicatively the belief and one in which the decay affected exponentially the belief. None provided significantly better fits than the equivalent model without decay. Furthermore, the decay parameter was very close to 1, meaning that adding a decay did not provide a better explanation of behavior. Both the Bayesian inference model with online learning and the Decay model poorly explained subjects' choices, and therefore will not be displayed (paired t-test against Standard RL, all $p < 10^{-5}$).

5.1.6.5 Mixed model

This model consisted of a mixture of the above RL and Bayesian systems, with a weight parameter ω arbitrating between them. It thus included more free parameters (learning rate, volatility, weight) than each system (RL or Bayesian) alone.

The relevant quantity for choice was thus:

$$\omega\ Q_t + (1 - \omega)\ log(p_t)$$

We allowed ω to vary across subjects and across sessions, but we assumed it to be constant throughout the experiment. ω could represent a measure of individual variability, according to participants' preferred reliance on Bayesian or RL system.

Repetition bias. We reproduced each of the above models with an additional parameter: a repetition bias modeling the tendency to stick with the previous action. For example, for reinforcement learning, after update, expected Q values for each action were simply modified according to:

$$\begin{cases} Q_{t,chosen} = Q_{t,chosen} + repetitionbias \\ Q_{t,unchosen} = Q_{t,unchosen} - repetitionbias, \end{cases} \tag{5.5}$$

This was motivated after observing that subjects switched less than what models would predict (see Results section).

5.1.6.6 Action selection

A general strategy for action selection was to stochastically select an action a_t according to the standard softmax rule (Luce, 1977 [104]), with parameters β and ϵ:

$$p(a_t = 1) = \frac{\epsilon}{2} + (1 - \epsilon)\frac{e^{\beta Val1_t}}{e^{\beta Val1_t} + e^{\beta Val2_t}}, \tag{5.6}$$

with $Val1_t$ and $Val2_t$ being the expected values for choosing option 1 and option 2 respectively, and β the softmax inverse temperature, allowing for exploration towards the lower-valued action. The optimal strategy is obtained when β tends towards infinity. In that case, the subject would pick, at each trial, the option with the largest expected value. The term with ϵ modeled the lapses proportion (e.g. trials with very short reaction times), with $\frac{1}{2}$ being the probability of random choice.

5.1.7 Fitting procedure

The fitting procedure objective was to find the set of free parameters that best fitted each subject's behavioral data. The number and nature of parameters depended on each particular model. For adjusting the models' free parameters to the behavioral data (subject' choices), we maximized the model log-likelihood (LLH):

$$LLH = \sum_t \log(p_t)$$

Where p_t was the probability that the model would have chosen the same action as the subject at trial t. Model fitting was done on all trials pooled from both sessions, which was justified given there was no evolution of behavior over the course of the experiment as shown in Figure 6.2. **Further details about fitting procedure are provided in the next Chapter.** Notably, we reproduced fits including only the second half of trials for each session, under the hypothesis that subjects had then reached a stable regime. This was especially critical for the Bayesian model, that included knowledge of reward distributions beforehand, which was unrealistic since subjects had no prior knowledge about the reward distributions and had to learn them by experience. However, they must have learnt them relatively rapidly because fits on the second half results were similar to what was obtained when including all trials, and the order of models in model comparison was not significantly changed.

5.1.7.1 Model selection

A crucial point in modeling is models comparison. Indeed, we could imagine a model that has a very high likelihood but that is not capturing well what subjects are doing. By contrast, a model with a large number of parameters could capture very well what subjects are doing, but is actually over-fitting the data (Hawkins et al., 2004 [168]). To prevent from this possibility, we evaluated both qualitative and quantitative measures for each model. We emphasize the importance of presenting qualitative models simulations to give an idea of the model behavior and to support its relevance for explaining the participants data.

5.1.7.2 Quantitative measures

The log-likelihood obtained for each fit gave an index of how well the model predicts the subject's choices. However, for a given model, the higher the number of free parameters added, the higher the log-likelihood, but it can be an artificial increase. To take into account the model complexity, we used the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC). These criteria take into account both goodness of fit and parsimony.

$$BIC = -\ 2\ LLH\ +\ k\ \ln(n)$$

$$AIC = 2\ k - 2\ LLH$$

with k the number of free parameters in the model. The BIC penalizes more the extra parameters because it accounts for the number of observations n (here, number of trials) used to fit the data. Quantitative measures can be misleading though, as log-likelihood overweighs low probability actions into the global calculation. Therefore, the following measures were also critical to examine whether the model qualitatively predicted the subject's choices.

5.1.7.3 Qualitative measures

Crucially, we assessed whether our models qualitatively reproduced subjects' behavior. To that aim, model simulations were performed. Taking the fitted parameters of each subject, the model was run as if it was a subject. It was then possible to study its choices sequence similarly as for participants (cf. Statistical Analyses). To assess whether model choices were consistent with subjects' choices, two behavioral measures were examined. We reproduced for each model's simulation the learning curves computed in Figure 6.1 as well as the stay/switch proportion following reception of each outcome as shown in Figure 6.3. Simulations provided a visual index of the model accuracy.

Another way of qualitatively comparing models was to look directly at fits instead of simulations. To do that, for each trial, the probability that the model would have made the same choice as the subject must be plotted. This measure was dependent on subjects' choices. Fits are not shown in this thesis because they were less sensitive than simulations in order to compare models.

5.2 Experiment 2

The second behavioral experiment was the same as the first one but instead of choosing between two stimuli, subjects had to choose between two tasks. We hypothesized that a higher level of abstraction (i.e. tasks instead of stimuli), subjects might rely on Bayesian inference rather than on reinforcement learning.

5.2.1 Experimental design

Stimuli were simple letters (e.g. A, e, n, D). At each trial, a stimulus appeared (Figure 5.6). Subjects were instructed to choose to perform either the discrimination task consonant/vowel or the discrimination task upper case/lower case, using four response keys (consonant, vowel, upper case, lower case).

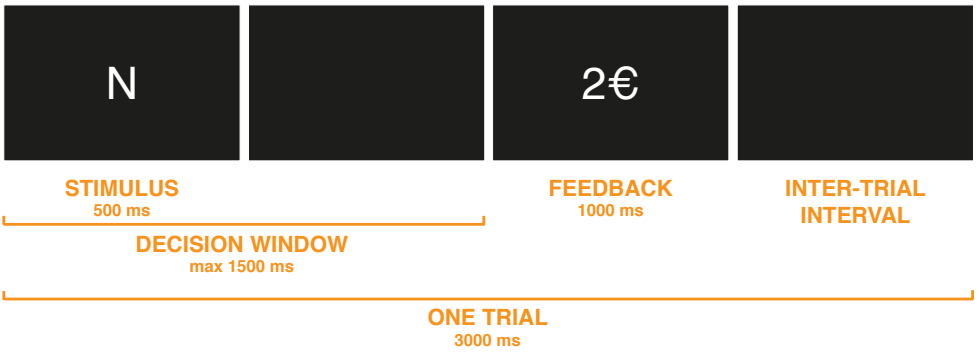


FIGURE 5.6: Trial structure and timing.

There was a highest rewarded task and a lower rewarded task (Figure 5.7), with the same reward distributions as Experiment 1. "Primes" at the beginning of each session were removed since they did not affect performance. So each subject did only one session for Experiment 2.

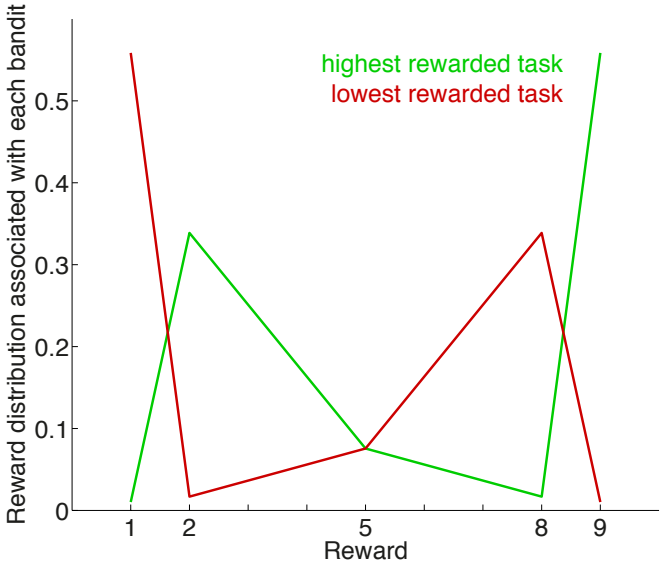


FIGURE 5.7: Reward distributions underlying each task.

Without any sensory cues, the highest and lowest rewarded tasks shifted after an unpredictable number of trials (reversals). Subjects had to learn which of the two tasks was the highest rewarded task, knowing that the best of the two tasks could change over time, and to respond in order to win as much as possible. Whatever the chosen task, in case the motor action was incorrect, the subject received 0 Euro (e.g. choice of the consonant/vowel task but response on the "consonant" key whereas the stimulus was a vowel).

Subjects underwent incremental training to gradually familiarize with the various experiment aspects. More precisely, they first trained separately on each task, doing the consonant/vowel categorization task and the upper case/lower case categorization task, receiving binary feedback (correct/incorrect). Then, they did a task version in which the choice between task was cued (the color of the letter stimulus indicated the task to perform), still receiving binary feedback (correct/incorrect). Therefore, subjects could handle the mapping between the four possible responses and the buttons. Finally, they trained on the actual experiment, freely choosing between the two tasks, and receiving parametric rewards drawn from 5.7.

5.2.2 Randomization, Experiment presentation and Participants

Each of the four types of stimuli (consonant upper case, consonant lower case, vowel upper case, vowel lower case letter) appeared an equal number of times in a random order. The four response keys location were counterbalanced across subjects. Critically, the highest rewarded distribution was associated to a task as many times as to the other task. We tested 24 new participants under the same recruitment conditions as in Experiment 1.

5.2.3 Statistical analysis and modeling

Statistical analysis and modeling were strictly similar to what was done for Experiment 1 with stimuli.

5.3 Experiment 3

The third task consisted of a control experiment to investigate whether the observed effects were dependent on the value scale we originally chose.

5.3.1 Experimental design

We modified the experiment value scale according to Figure 5.8. The reward distribution remained with the same probabilities, but with different outcome values: 1, 3, 5, 7 and 9 euros.

The Experiment 1 paradigm was re-used, i.e. choice between two stimuli. The “primes” at beginning of session were kept for consistency with Experiment 1, even if they provided

no interesting effect. So each subject did two sessions, as in Experiment 1. Therefore we were able to compare Experiment 1 and Experiment 3 modifying only the value scale, everything else being equal.

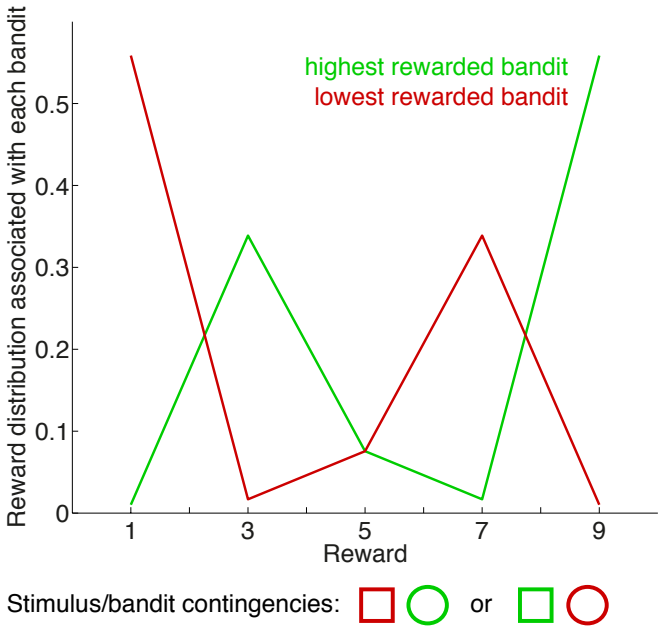


FIGURE 5.8: Reward distributions underlying each bandit, with modified value scale.

The trial structure and timing was similar to that of Experiment 1 (Figure 5.9).

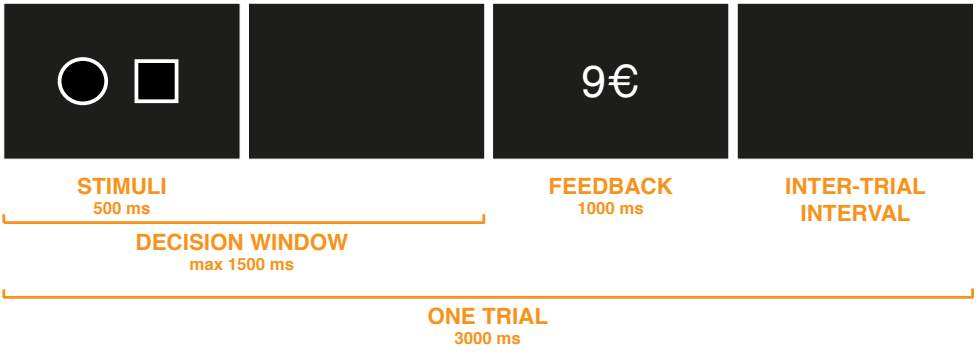


FIGURE 5.9: Trial structure and timing.

We tested 13 new participants under the same recruitment conditions as in Experiment 1.

5.3.2 Statistical analysis and modeling

Statistical analysis and modeling were strictly similar to what was done for Experiment 1 with stimuli.

Chapter 6

Protocol A: Results and Discussion

We present the behavioral and modeling results of a series of three probabilistic reversal-learning tasks, involving stochastic and changing reward structures, aiming at dissociating affective from informational values of action outcomes.

6.1 Experiment 1

In this first experiment, we investigated whether subjects use affective and/or belief values of rewards to guide their decisions.

6.1.1 Experiment 1: Behavioral Results

Figure 6.1 represents the choice proportion of the highest rewarded bandit plotted against time after a reversal, averaged over 23 subjects.

Learning curves in Figure 6.1 showed that after a reversal, subjects were able to learn to choose the highest rewarded bandit (all trials, both sessions pooled). After 5-10 trials following a reversal, they reached an asymptotic level around 75%. On average, they chose the highest rewarded of the two bandits in 69.4% of trials.

In addition, no progression over the course of sessions was observed. An ANOVA with factor EPISODE NUMBER and subjects as repeated measures revealed no significant effect of episode number ($p = 0.63$). Indeed, Figure 6.2 shows that subjects did not tend to choose more and more the highest rewarded bandit as the session progressed.

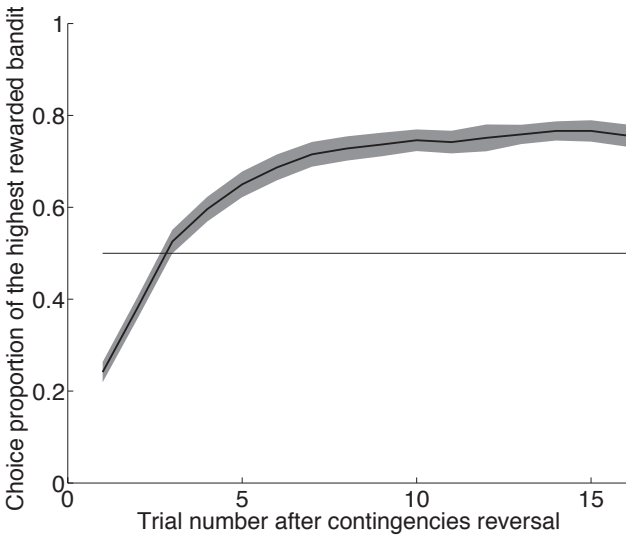


FIGURE 6.1: Experiment 1: Learning curves representing choice proportion of the highest rewarded bandit after a contingencies reversal. Subjects' behavior (N = 23) is displayed in black, with shaded area representing the standard error of the mean. The horizontal line represents chance level (50%).

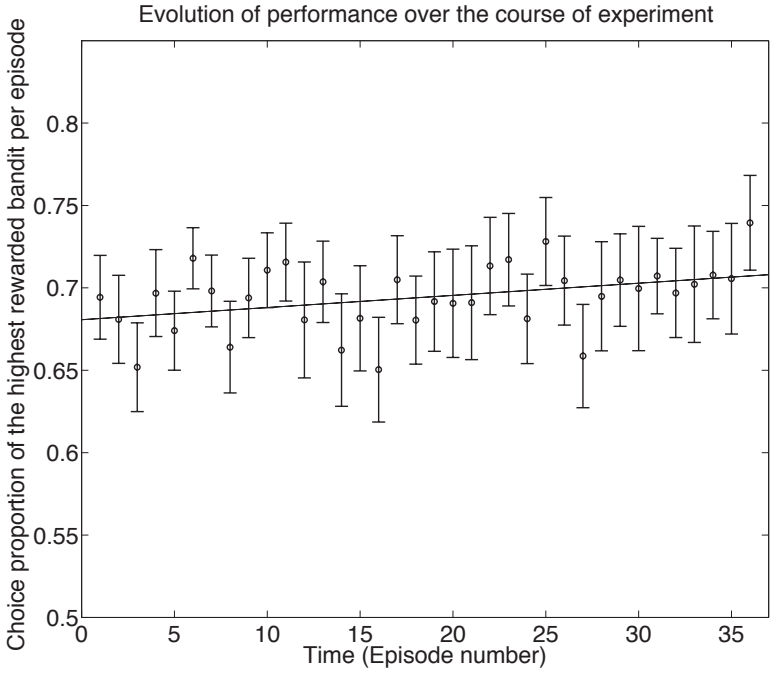


FIGURE 6.2: Experiment 1: Choice proportion of the highest rewarded bandit per episode, showing subjects' progression throughout the session (averaged across both sessions). Error bars: standard error of the mean, N = 23 subjects.

We observed a slightly lower general performance when subjects started with the session with "Bayesian prime". However, we observed an important inter-individual variability, both in the fits and in the behavior. Compared with previous experiments conducted within the team, informal debriefing revealed that subjects seemed very troubled with the experiment, reporting for example: "Was it possible to learn the best option or was it just random?" or "I thought the best shape changed every 5 trials" or "I was not able to identify any logic". We interprete this confusion in relation to the counter-intuitive bimodal reward distributions. Note: no subject was excluded on the basis of the informal debriefing. Only one subject was excluded for performing at chance level.

Additionally, we examined in Figure 6.3 the proportion of trials in which subjects stayed on the same choice according to the outcome received at previous trials. More precisely, we hypothesized that computing the switch/stay behavior after a reward of 8 Euros (and, symmetrically, of 2 Euros) would allow us to discriminate between subjects who did or did not infer the underlying task structure (cf. reasoning in Experiment 1 Methods). If subjects were only sensitive to reward magnitudes, we should have observed that the higher the obtained reward, the higher the stay proportion, the more the subject would repeat the same choice on the next trial. By contrast, if subjects had inferred the reward distributions underlying each bandit, they would have switched more after reception of 1 or 8 Euros than after reception of 2 or 9 Euros, as predicted by a Bayesian model (Figure 6.4). We observed that subjects did not switch more after gain of 8 Euros compared to gain of 2 or 5 Euros, as revealed in Figure 6.3. Consistently, the lowest stay proportion was observed after reception of the lowest outcome (1 Euro), while the highest stay proportion was observed after reception of the highest outcome (9 Euros).

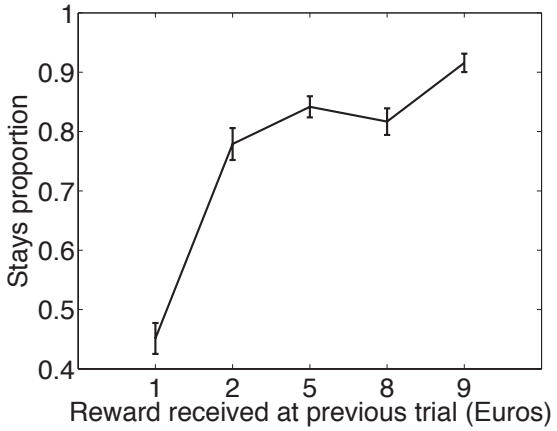


FIGURE 6.3: Experiment 1: Stay/switch trials proportion given reward received. Error bars: standard error of the mean, N = 23 subjects.

This behavioral pattern was rather consistent with a RL model. To further examine the cognitive mechanisms underlying the observed behavior, we examined several computational models.

6.1.2 Experiment 1: Modeling Results

Generally, average subjects' behavior lay between RL and Bayesian simulations as shown in Figure 6.4. Learning curves simulations showed that all models were able to re-learn, after a reversal, to choose the highest rewarded bandit. Both standard and normalized RL models adapted slower when a reversal occurred (red and brown curves in left panel Figure 6.4), and were slower to reach the asymptote. As predicted, Bayesian model simulations showed a significantly higher propensity to switch after receiving 8 Euros than subjects, since the model inferred that 8 Euros was likely to be a hallmark of the lowest rewarded bandit (blue curve in right panel in Figure 6.4).

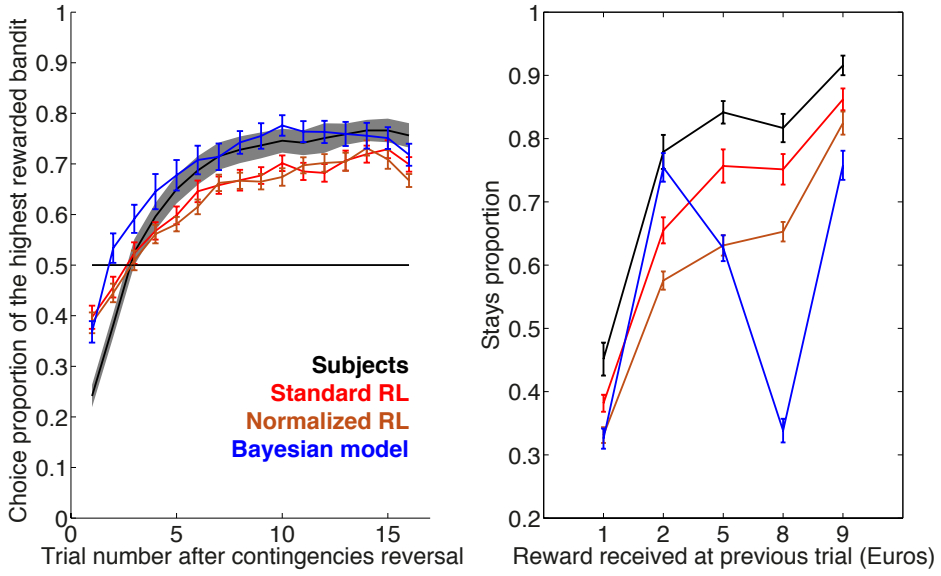


FIGURE 6.4: Experiment 1: Learning curves representing choice proportion of the highest rewarded bandit after a contingencies reversal (left panel) and stay/switch trials proportion given reward received (right panel). Subjects' behavior is displayed in black and models' simulations are displayed in color. The horizontal line in left panel represents chance level (50%). Error bars and shaded area represent the standard error of the mean, N = 23 subjects.

The slight drop in *Stays proportion* after receiving 8 Euros compared to after receiving 5 Euros was reproduced in Standard RL simulations (Figure 6.4). However, none of the models alone was able to qualitatively capture subjects' behavior.

The mixed model better captured subjects' behavior than either RL or Bayesian model separately, both qualitatively as shown in simulations in Figure 6.5 and quantitatively in model comparison (Figure 6.6). The mixed model (best fit) explained significantly better the behavior than the Standard RL model (second-best fit) (paired t-test on LLH: $p < 0.01$, on BIC: $p < 0.02$, and on AIC: $p < 0.01$), despite having two additional free parameters compared to the Standard RL model.

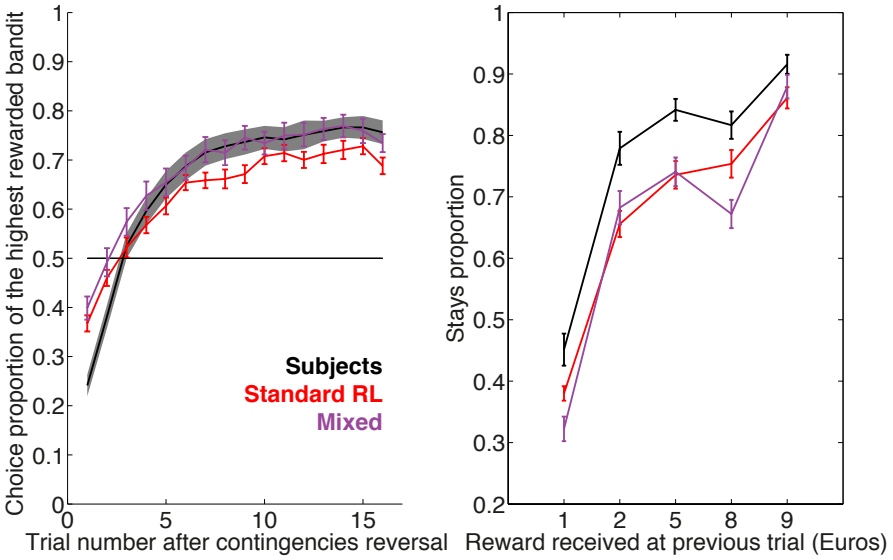


FIGURE 6.5: Experiment 1: Learning curves representing choice proportion of the highest rewarded bandit after a contingencies reversal (left panel) and stay/switch trials proportion given reward received (right panel). Subjects' behavior is displayed in black and models' simulations are displayed in color. The horizontal line in left panel represents chance level (50%). Error bars and shaded are represent the standard error of the mean, N = 23 subjects.

Nevertheless, the mixed model had a different behavior just after a reversal (left panel in Figure 6.5) and the pattern for stay/switch given reward received at previous trial was also different (right panel in Figure 6.5). The best-fitting mixed model parameters are provided in Table 6.1.

Parameters	**Description**	**Mean**	**S.E.M.**
β	Inverse temperature in softmax	10.3	6.0
volatility	Volatility in Bayesian system	0.20	0.05
ϵ	Lapses rate in softmax	0.11	0.02
learning rate	Learning rate in RL system	0.74	0.04
ω	Weight between the two systems in decision	0.55	0.07

TABLE 6.1: Best-fitting mixed model parameters. Mean and standard error of the mean (S.E.M.) across subjects (N = 23) are provided.

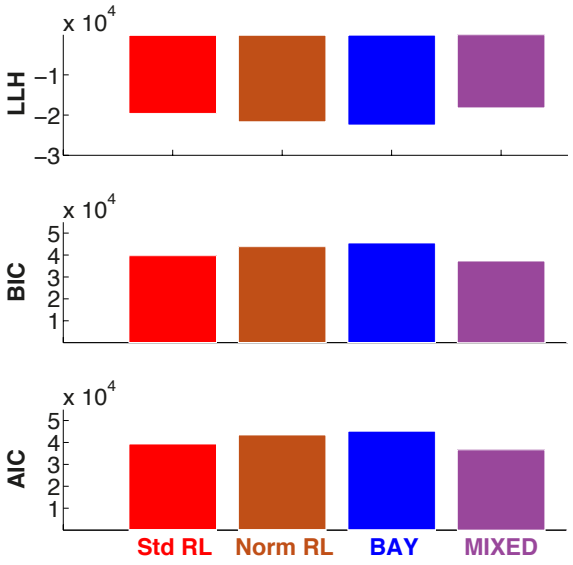


FIGURE 6.6: Experiment 1: Models selection, fixed effects analysis. LLH, BIC and AIC, summed across subjects (N = 23) are presented for each model: Standard RL (red), Normalized RL (brown), Bayesian (blue) and Mixed (purple).

In particular, volatility was overestimated (average: 0.20) compared to its real value (reversal frequency: 0.04). Consistently, the fitted learning rate was relatively high (mean value across subjects: 0.74). The distribution of the weight parameter ω mixing RL and Bayesian did not have a clear mode (Figure 6.7). It means that ω might not have a relevant meaning regarding each system contribution to decision (RL/Bayesian), but probably just reflected different types of subjects.

Repetition bias. Since the subjects overall stayed more than the models, we added a repetition bias modeling the tendency to repeat previous choice ("stickiness"). We can then better reproduce the stays proportion after each reward received (right panel, Figure 6.8). The mixed model qualitatively fits both the subjects' learning curves and stays proportion. With the repetition bias, the mixed model still fitted better quantitatively the data than the Standard RL (paired t-test on BIC: $p < 0.05$). However, simply adding a repetition bias is not very satisfactory in terms of explanatory power.

Lastly, the Bayesian model that learnt online reward distributions along with trial-by-trial learning of the stimulus/bandit contingencies did not account for subjects' behavior, neither in terms of qualitative simulations nor of quantitative criteria (paired t-tests on LLH, BIC and AIC compared to all other models: all $p < 10^{-5}$). A possible explanation for this poor result would be the model's complexity, with a lot of variables to monitor, probably associated with a high cognitive load.

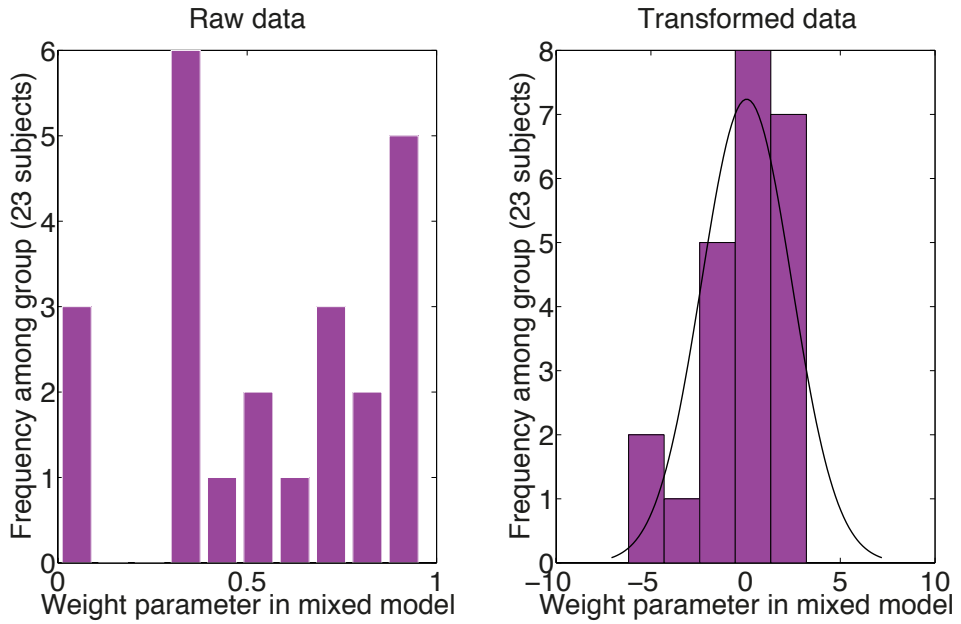


FIGURE 6.7: Experiment 1: Distribution of the fitted weight parameter within the group. Left panel represents raw data in the form of ω i.e. the contribution of the RL system. Right panel: same data after log transformation, with a gaussian fit.

6.1.3 Experiment 1: Discussion

In terms of quantitative criteria, the best fit at the group-level was the mixed model. The mixed model provided slightly better results but we think it did not considerably improve our understanding of subjects' behavior in the task, compared to what predicted the Standard RL model. A few subjects' behavior remained unexplained. The behavior of these few subjects was much more consistent with the Bayesian model than with the RL. However, it could be that these few participants simply learned a heuristic rule, noticing the association 8/1 Euros with the lowest bandit vs. 9/2 Euros with the highest bandit, and apply it efficiently. These particular subjects' behavior was not satisfactorily explained, but need to be further looked in the light of the next series of experiments.

Fitted volatility in the Bayesian and in the Mixed model was overestimated compared to its real value. This result has been consistently observed across different paradigms in our team. A possible explanation could be that this parameter would be a second order estimate. Consequently, subjects had more difficulty estimating it, or perceived the environment as more volatile or more uncertain than it was in reality. In our task, there were two main sources of uncertainty. On the one hand, expected uncertainty, as termed by Bossaerts and colleagues [144] [160], relates to noisy feedback even though the subject chose the highest rewarded bandit. On the other hand, unexpected uncertainty relates to the presence of contingencies reversal, that subjects could not anticipate.

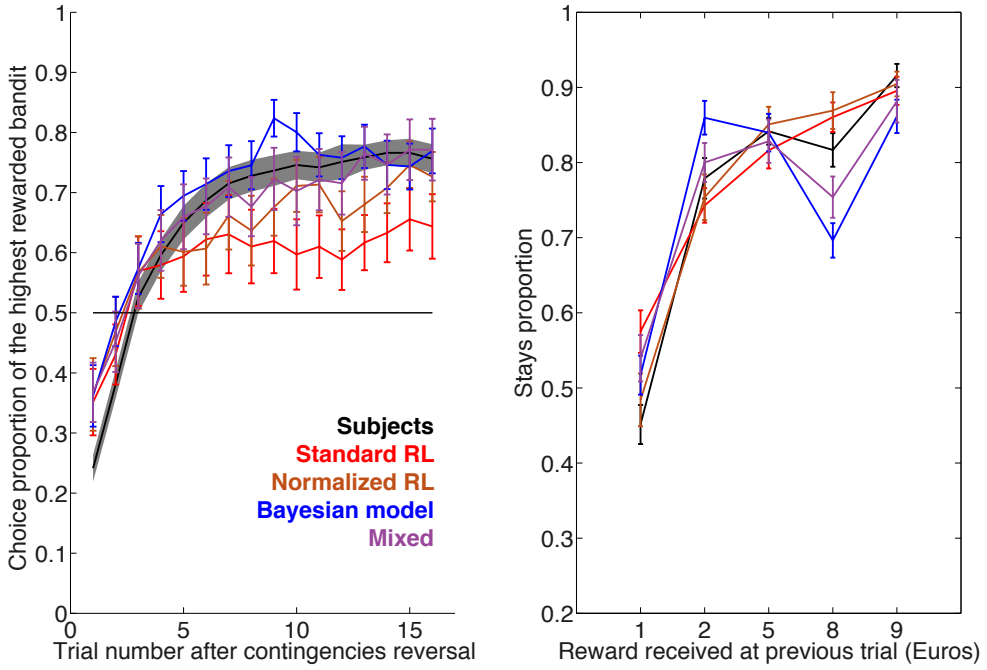


FIGURE 6.8: Experiment 1: Learning curves representing choice proportion of the highest rewarded bandit after a contingencies reversal (left panel) and stay/switch trials proportion given reward received (right panel). Subjects' behavior is displayed in black and models' simulations are displayed in color: Standard RL (red), Normalized RL (brown), Bayesian (blue) and Mixed (purple). All models included a repetition bias modeling the tendency to reproduce previous choice. The horizontal line in left panel represents chance level (50%). Error bars and shaded are represent the standard error of the mean, N = 23 subjects.

More broadly, it seems that subjects did not deal with a continuous reward scale, but transform each parametric outcome into a binary outcome i.e. map rewards onto a dichotomous feedback scale, such as "success/failure" or "good/bad". Indeed, in our task, the brain seemed unable to manage a distribution in which reward probability was not correlated with reward magnitude. A possible working hypothesis is that the brain would include an interface transforming continuous parametric rewards into binary outcomes, with a threshold to be defined, and then would decide to stay or switch options given the obtained reward was below of above the threshold. This dichotomous hypothesis differs from a RL in the sense it accounts for the notion of switch, which is a discrete event, whereas updates in a RL are continuous. This hypothesis of reward binarizing will not be investigated further in this thesis, but possible neural systems implementing such a dichotomous model could involve the basal ganglia, being implicated in value-based learning, whereas the prefrontal cortex could perform Bayesian inference on binary states extracted from continuous values. vmPFC could be a candidate for being at the interface.

Finally, the predominance of affective values in decision in our task, rather than beliefs about states, might be explained from an evolutionary point of view. Perhaps the brain architecture has been selected for strongly attributing values to stimuli or objects. So it could be that we would obtain different results for a choice between motor actions or a choice between tasks or task-sets, as compared to a choice between stimuli. We further investigated this hypothesis in Experiment 2, in which choice between tasks replaced choice between stimuli. We hypothesized that at a higher hierarchical level of representations, a higher degree of abstraction (i.e. tasks instead of stimuli), affective values might be less salient and subjects would be less sensitive to pure reward magnitude. We expected this second experiment to favor the emergence of Bayesian learning.

6.2 Experiment 2

The second behavioral experiment was the same as the first one but instead of choosing between two stimuli, subjects had to choose between two tasks. We hypothesized that at a higher level of abstraction, i.e. tasks instead of stimuli, subjects might rely more on inferential systems than on pure reinforcement learning.

6.2.1 Experiment 2: Behavioral Results

Unexpectedly, behavior was very similar to what was observed for a choice between stimuli in Experiment 1. Learning curves in Figure 6.9 showed that after a reversal subjects were able to adapt and learn to choose the highest rewarded task, eventually reaching a plateau around 75%. General performance was slightly lower than in Experiment 1, with 5 subjects out of 24 performing at chance level, that were removed from the following analyses. Despite training, the experiment was highly demanding in terms of cognitive load (learning of 2 tasks, 4 response buttons, mapping of tasks onto reward distributions ...).

A slight progression of performance over the course of episodes was observed in Experiment 2 (Figure 6.10). An ANOVA with factor EPISODE NUMBER and subjects as repeated measures revealed a trend effect of episode number ($p = 0.06$). However, subjects remained less performant overall than in Experiment 1. Indeed, on average, they chose the highest rewarded of the two bandits in 65.8% of trials.

As found in Experiment 1, the lower the reward received, the more the subjects switched on the following trial (Figure 6.11). Notably, no significant difference in switch proportion was observed after reception of 5, 8 or 9 Euros. However, the average stay

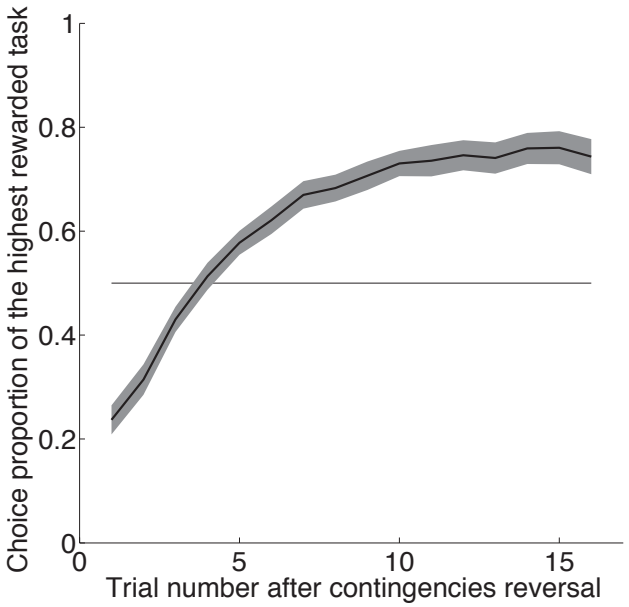


FIGURE 6.9: Experiment 2: Learning curves representing choice proportion of the highest rewarded task after a contingencies reversal. Subjects' behavior is displayed in black and models' simulations are displayed in color. The horizontal line represents chance level (50%). Shaded area: standard error of the mean, N = 19 subjects.

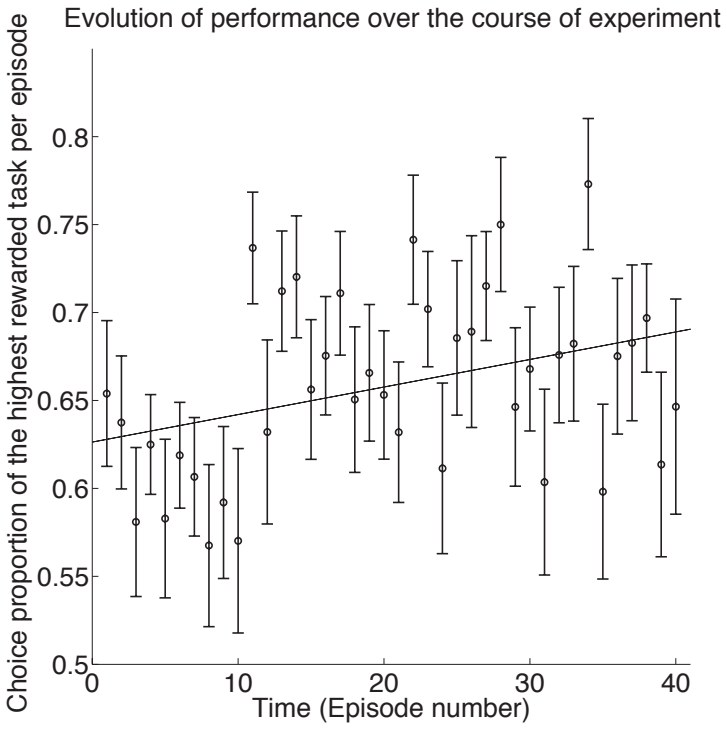


FIGURE 6.10: Experiment 2: Evolution of the choice proportion of the highest rewarded task over the course of the experiment. Error bars: standard error of the mean, N = 19 subjects.

proportion was not significantly different after reception of 2, 5, 8 and 9 Euros, but was significantly higher than after reception of 1 Euro. This dichotomy supports the hypothesis of outcomes "binarization" proposed in Experiment 1 discussion.

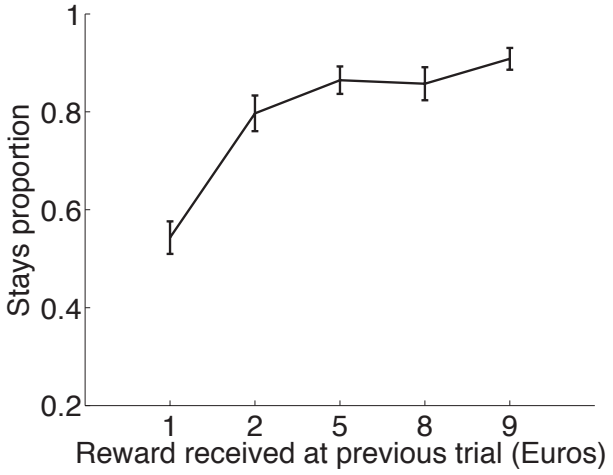


FIGURE 6.11: Experiment 2: Stay/switch trials proportion given reward received. Error bars: standard error of the mean, N = 19 subjects.

These behavioral measures will be used to compare models in the next section.

6.2.2 Experiment 2: Modeling Results

Modeling results generally replicated what was observed for a choice between two stimuli in Experiment 1. Models simulations presented in Figure 6.12 indicated that qualitatively, all Standard RL, Normalized RL, Bayesian, and Mixed model captured the subjects' learning curves (left panel), especially just after a reversal. However, subjects stayed on the same choice more than what models predicted (right panel), meaning that neither model could not be totally accurate. In particular, subjects were far from the Bayesian model predictions.

Nevertheless, the mixed model provided the best fit in terms of quantitative criteria (Figure 6.13). In particular, the mixed model better fitted the behavior than all other models in terms of LLH and AIC (paired t-tests, all $p < 0.02$). However, as opposed to Experiment 1, the mixed model was not significantly better than the second-best fitting model (Standard RL) in terms of BIC (paired t-test, $p = 0.27$). The BIC is a more conservative criterion than AIC (cf. Methods).

In summary, Experiment 2 results were similar to Experiment 1 although Experiment 2 was slightly more difficult.

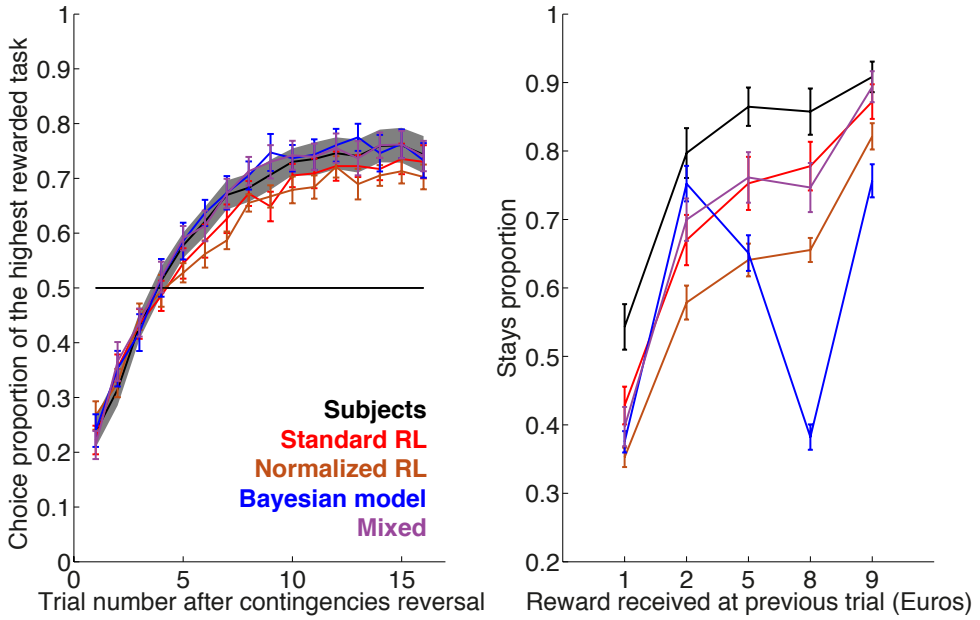


FIGURE 6.12: Experiment 2: Learning curves representing choice proportion of the highest rewarded task after a contingencies reversal (left panel) and stay/switch trials proportion given reward received (right panel). Subjects' behavior is displayed in black and models' simulations are in color. The horizontal line in left panel represents chance level (50%). Error bars: standard error of the mean, N = 19 subjects.

6.2.3 Experiment 2: Discussion

Experiment 2 data did not support our original hypothesis i.e. the level of tasks instead of stimuli, more abstract, would favor the emergence of Bayesian inference. It seems that the brain had much difficulty dealing with bimodal distributions in which affective values and belief values were not correlated.

None of the tested computational models revealed a satisfactory description of behavior, although the standard RL provided the most convincing and parsimonious fit. It could be that when the task is difficult and demanding in terms of cognitive load, subjects go back to simpler and robust behavior such as RL.

6.3 Experiment 3

In this third experiment, we investigated whether the observed effects were dependent on the particular reward scale that we used in Experiment 1.

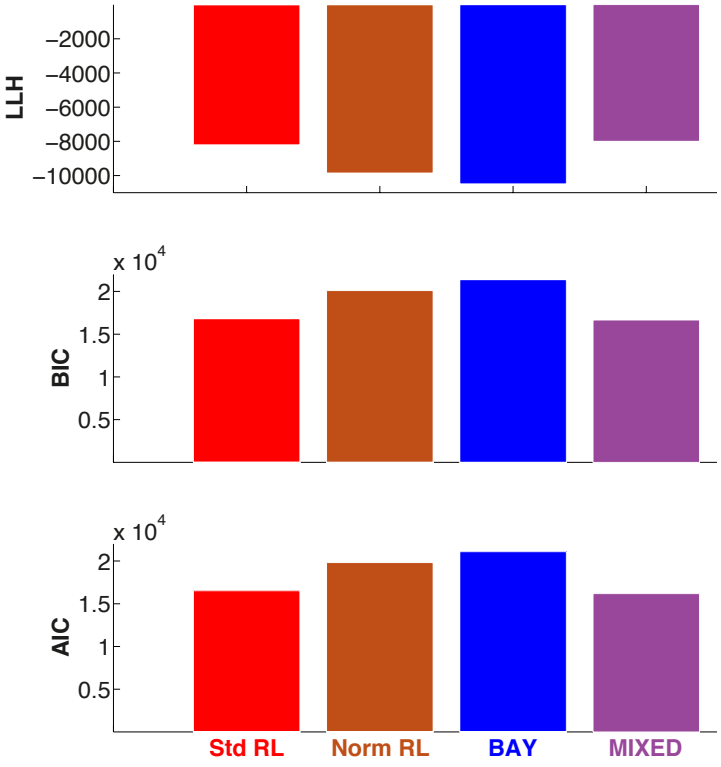


FIGURE 6.13: Experiment 2: Models selection. LLH, BIC and AIC, summed across subjects (N = 19) are presented for each model: Standard RL (red), Normalized RL (brown), Bayesian (blue) and Mixed (purple).

6.3.1 Experiment 3: Behavioral Results

We observed a similar behavior as in Experiment 1, with the same qualitative pattern for learning curves (Figure 6.14). After a reversal, subjects were able to re-learn which bandit was the highest rewarded one, and reached an asymptotic level above 80%. On average, they chose the highest rewarded of the two bandits in 74.6% of trials, thus better performing on this more balanced reward scale (1-3-5-7-9 Euros) than on Experiment 1 reward scale (1-2-5-8-9 Euros).

No progression of performance over the course of episodes was observed in Experiment 3 (Figure 6.15). An ANOVA with factor EPISODE NUMBER and subjects as repeated measures revealed no significant effect of episode number ($p = 0.19$).

We then examined the stay proportion (stick with the same choice) as a function of the reward received at previous trial (Figure 6.16). This behavioral measure revealed that the higher the reward received, the more the subjects persevered with the same stimulus on the next trial, as observed in Experiment 1.

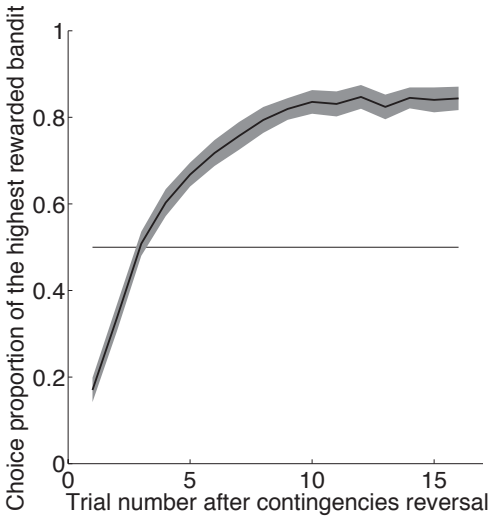


FIGURE 6.14: Experiment 3: Learning curves representing choice proportion of the highest rewarded bandit after a contingencies reversal. Subjects' behavior is displayed in black and models' simulations are displayed in color. The horizontal line represents chance level (50%). Shaded area: standard error of the mean, N = 12 subjects.

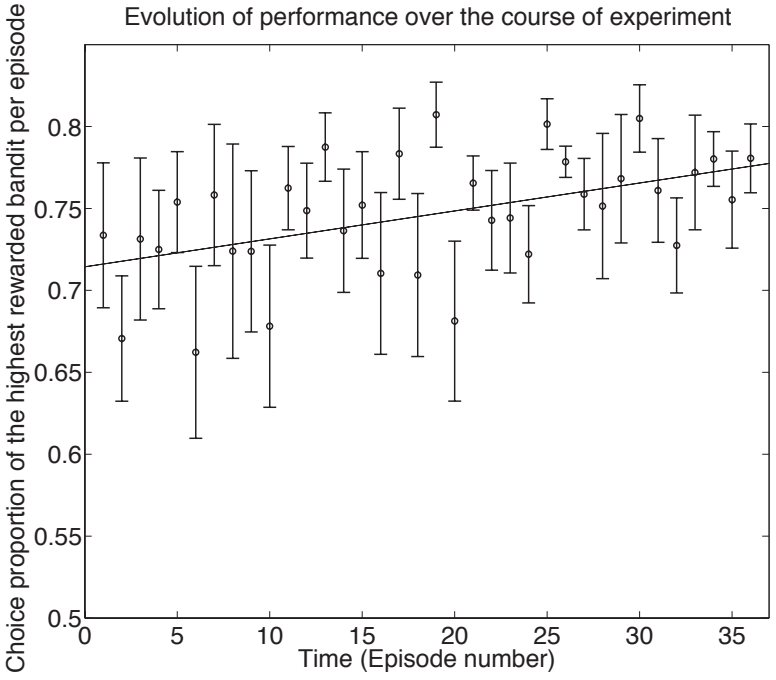
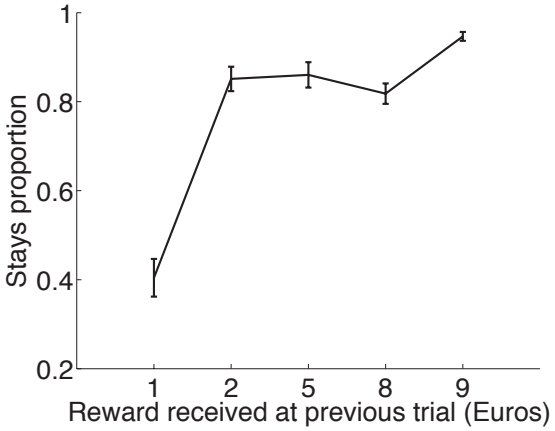


FIGURE 6.15: Experiment 3: Evolution of the choice proportion of the highest rewarded bandit over the course of the experiment (averaged over both sessions). Error bars: standard error of the mean, N = 12 subjects.



proportion

FIGURE 6.16: Experiment 3: Stay/switch trials proportion given reward received. Error bars: standard error of the mean, N = 12 subjects.

6.3.2 Experiment 3: Modeling Results

Models simulations provided the same qualitative pattern for each model as for Experiments 1 (Figure 6.17). In particular, RL models adapted slower when a reversal occurred.

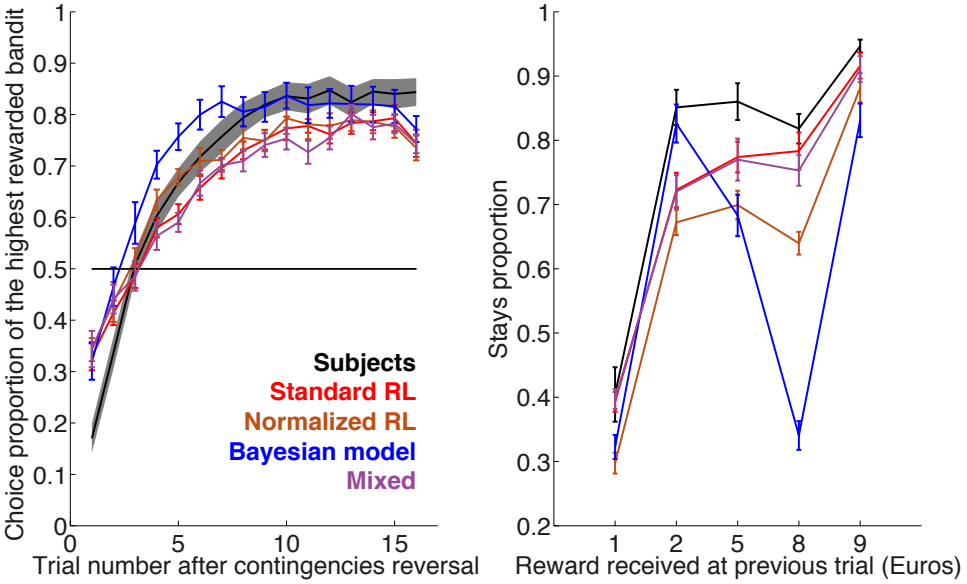


FIGURE 6.17: Experiment 3: Learning curves representing choice proportion of the highest rewarded bandit after a contingencies reversal (left panel) and stay/switch trials proportion given reward received (right panel). Subjects' behavior is displayed in black and models' simulations are displayed in color. The horizontal line in left panel represents chance level (50%). Error bars: standard error of the mean, N = 12 subjects.

Moreover, quantitative criterion (i.e. LLH, BIC and AIC) led to similar model selection as for Experiment 1, with the mixed model better explaining data than any other alternative models (all $p < 0.005$, Figure 6.18).

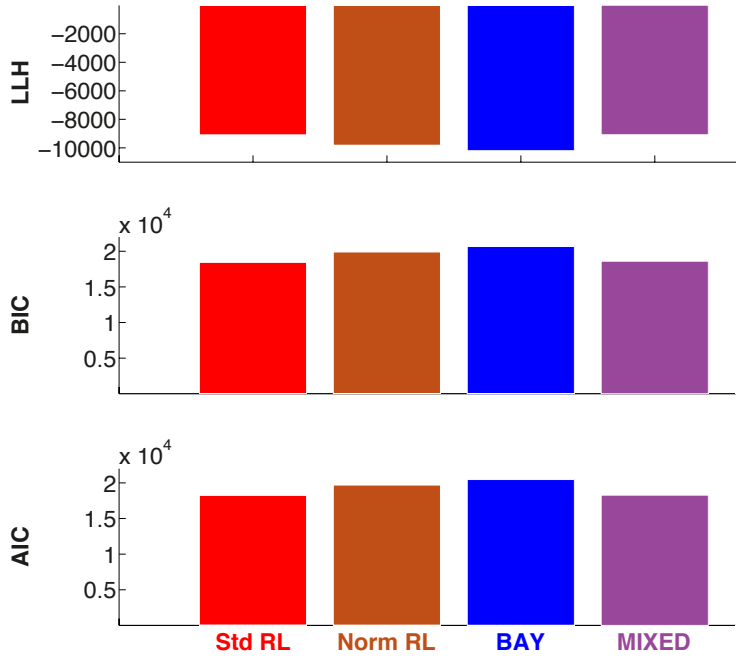


FIGURE 6.18: Experiment 3: Models selection. LLH, BIC and AIC, summed across subjects (N = 12) are presented for each model: Standard RL (red), Normalized RL (brown), Bayesian (blue) and Mixed (purple).

However, as in Experiments 1 and 2, models simulations presented in Figure 6.17 showed that neither Standard RL, Normalized RL, Bayesian, nor Mixed model fully captured the subjects' behavior.

6.3.3 Experiment 3: Discussion

In Experiment 3, we replicated the results of Experiment 1 using different reward distributions. We concluded from Experiment 3 that the particular value scale neither impacted qualitatively the behavior nor quantitatively the model selection.

6.4 Experiments 1, 2 and 3: Conclusion

Taken together, these results support the hypothesis of a psychological process converting continuous outcomes into binary states, as proposed in Experiment 1 Discussion.

Subjects do not seem able to detect the task structure i.e. the bimodal reward distributions, and were unable to use it to respond efficiently.

Grouping these three behavioral experiments' results, the mixed model remains globally the best-fitting model. However, it does not provide a much more satisfactory explanation of behavior than a standard reinforcement learning model, simpler and more parsimonious. In other words, the discrimination between the mixed model and the Standard RL remained unclear. Therefore, to further investigate whether the mixed model could truly be a reliable explanation of subjects' choices, we next developed another paradigm.

Chapter 7

Protocol B: Integration of beliefs and affective values in decision-making

7.1 Probabilistic reversal-learning task

7.1.1 Paradigm

The task aims at separating two conceptual dimensions of rewards: the affective value and the informational value. The affective value refers to the subjective value, experienced over a continuous axis of subjective preferences. By contrast, the informational value is a more abstract concept, and refers to the information carried by the reward about choice's reliability. We addressed the question of how informational and affective values are integrated, and of how information-carrying values influence the subject's behavior.

Subjects carried out a decision-making task, repeatedly choosing between two stimuli (two shapes: a square and a diamond). The potential rewards to be possibly won were displayed in the centre of the shapes before each choice (Figure 7.1; shapes are colored for clarity but were white during the actual experiment).

Subjects were instructed to maximize gains, knowing that:

- One of the two shapes led to obtain a reward more frequently than the other one;

- The shape that most frequently led to obtain the proposed reward could reverse from time to time. Reversals structure is displayed in Figure 7.2.

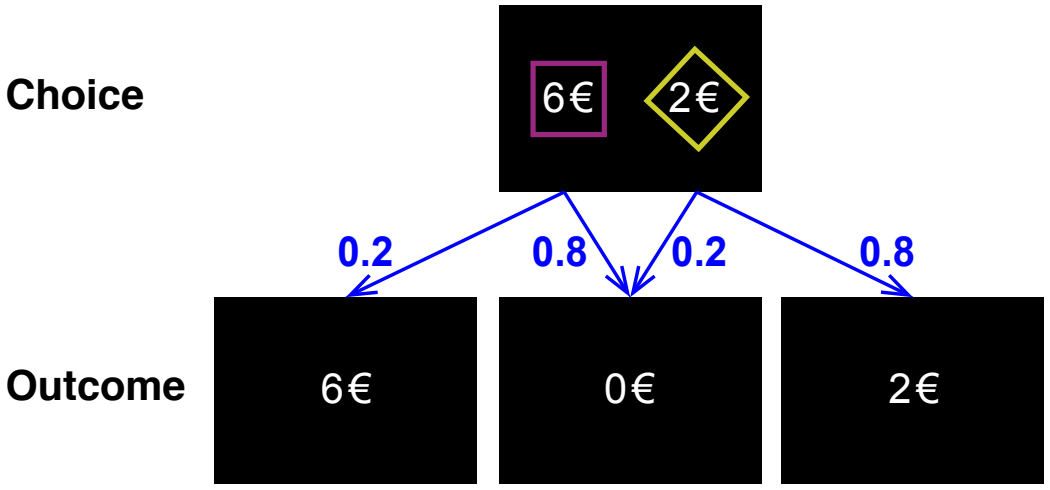


FIGURE 7.1: Probabilistic reversal learning task: Trial structure.

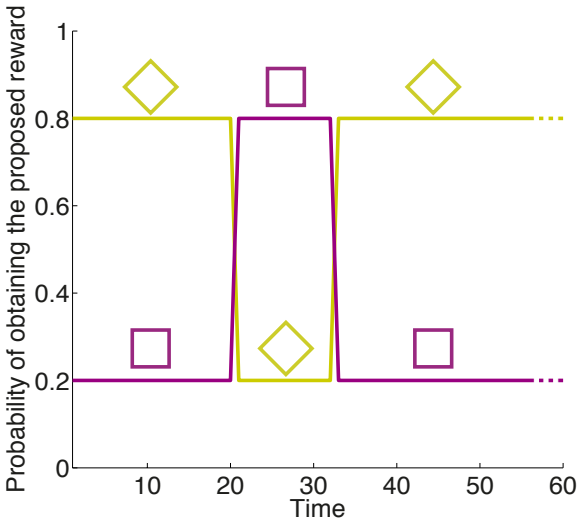


FIGURE 7.2: Probabilistic reversal learning task: Reversals structure.

The reward could be either received ($x_t = 1$, with probability 0.8), or not received ($x_t = 0$, with probability 0.2). In other words, each shape could have two possible states: frequently rewarded (with probability 0.8) or not frequently rewarded (with probability 0.2) (see Generative model of the task in Appendix D). Additionally, in case the reward was received ($x_t = 1$, with probability 0.8), it could be either the proposed reward (r_t with probability 0.5) or a close one ($r_t + 1$ with probability 0.25 or $r_t - 1$ with probability 0.25) (not shown in Figure 7.1). Based on the results of two behavioral pilot tasks, this possible small discrepancy between the reward proposed and the reward received was introduced in order to have a good balance between the use of rewarding

values and informational values by subjects. Otherwise, the outcome would have been perceived simply binary (gain or loss of the proposed reward), providing no emphasis on parametric affective value.

Crucially, we manipulated the reward distributions underlying each state, in order to modulate the link between proposed rewards and states. Proposed rewards associated with each shape were drawn among five possible values: 2, 4, 6, 8 and 10 Euros. Three experimental conditions were consequently established (Figure 7.3):

- **Correlated Values condition** Higher proposed rewards were correlated with the most frequently rewarded state (0.8). In this condition, it meant that choosing higher proposed rewards was better on average. In other words, higher rewards occurred more in the most frequently rewarded state. Proposed rewards were drawn from an exponential-like distribution with $\gamma = 0.13$ (see Generative model of the task in Appendix D).
- **Random Values condition** For each shape, rewards were randomly drawn among the five possible values ($\gamma = 0$, flat distribution). This consisted of a baseline condition. As rewards were randomly drawn, they conveyed no information about the underlying state to which the shapes were associated. In other words, proposed rewards carried no cue about which shape was the most frequently rewarded state. The Random Values condition actually corresponds to the task presented by Behrens and colleagues, 2007 [162].
- **Anti-correlated Values condition** Conversely to the Correlated Values condition, higher rewards were correlated with the least frequently rewarded state (0.8). It meant that higher rewards occurred more often with the least frequently rewarded state. In other words, lower rewards occurred more often in the most frequently rewarded state. Rewards were drawn from an exponential-like distribution with $\gamma = -0.13$. In this condition, there was thus a conflict between proposed rewards and states.

In the Correlated Values condition alone, it was not possible to tell apart both affective and informational values, since both varied in the same direction. On a given trial, both proposed rewards were drawn separately for each shape. Therefore, there was no statistical link between the two proposed rewards on a given trial. Moreover, the exponential reward distribution slope, γ, was set such that:

1) The most and least frequently rewarded options were not flipped in the anti-correlated condition. More precisely, the state with the global highest expected value corresponded to the most frequently rewarded state on average. It means that, on average, it was still

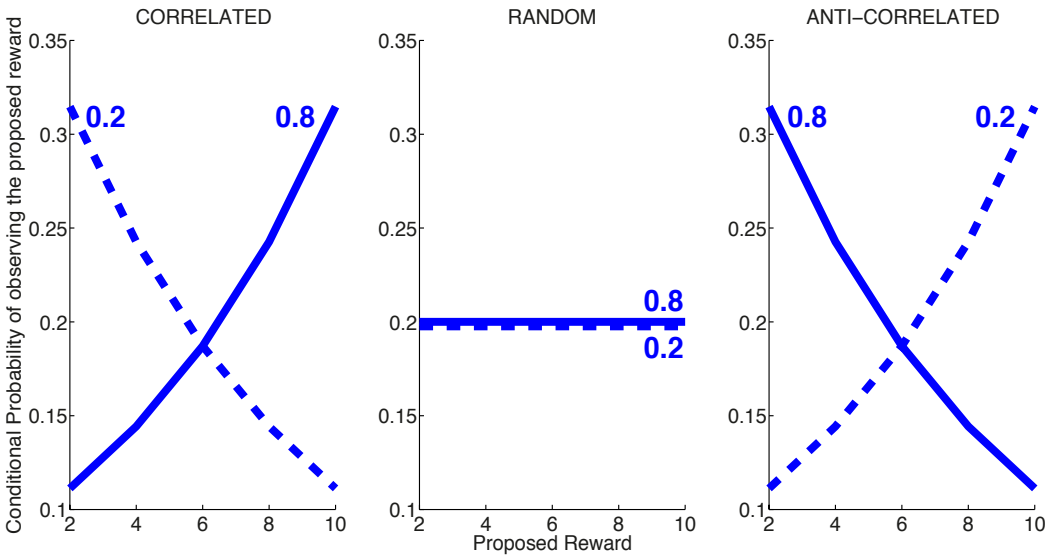


FIGURE 7.3: Probabilistic reversal learning task: Reward distributions for the three experimental conditions: correlated values, random values, anti-correlated values.

worth to choose the most frequently rewarded state (the one with probability 0.8 of leading to a reward), even though it was proposing lower rewards. In other words, in all conditions, and in particular in the anti-correlated values condition, the Pascalian utility was still maximum for the most frequently rewarded state.

2) Rewards with low probability of occurrence (at one extreme of the distribution) were not too low so that they still regularly appeared.

Two pilot studies were conducted before this protocol was run under fMRI (respectively 25 subjects and 12 subjects), which allowed to adjust the task design. Different participants underwent the two pilot experiments and the fMRI experiment to avoid meta-learning.

All stimuli and feedbacks were presented using PsychToolBox (Brainard, 1997 [167]) and appeared on a uniform black background (Figure 7.1). Subjects underwent 400 trials per session. Each session was divided in four runs, roughly equally long (see Breaks below).

7.1.2 Design and Randomization

The task was carefully designed so that there were no low-level biases in the results. Otherwise, subjects tend to infer patterns when there are regularities or sequences in

the local history of trials (Gaissmaier and Schooler, 2008 [169]; Yu and Cohen, 2009 [145]).

Episodes

A sequence of trials in between two reversals is called an episode. Each session consisted of 20 episodes so 19 reversal events. Episodes were either 16, 20, 24 or 28 trials long. Episodes order was pseudo-randomized on half a session: 16, 16, 16, 16, 20, 20, 20, 24, 24, 28, with each quarter of session including exactly the same number of trials. Overall, four possible episodes lengths sequences (sequences A, B, C, D, see Table 7.1) were used. Within subject, the episodes lengths sequence was different for each condition. Episodes lengths sequences were co-controlled with condition and with execution order of conditions.

Stimuli

According to the design, when the most frequently rewarded shape was chosen, a reward was given 80% of the time. When the least frequently rewarded shape was chosen, a reward was given 20% of the time. For each shape, the probability of obtaining a reward switched between 0.8 (most frequently rewarded shape) and 0.2 every episode. Within each episode, the proportion 0.8/0.2 for each shape was controlled, but because episodes had small finite lengths, approximations were made when necessary and fully counterbalanced across episodes. Within each session, all first and second trials following a reversal were controlled for respecting the 0.8/0.2 balance too. The shape that started the session as being the most frequently rewarded shape was counterbalanced between conditions and between subjects. Nature and position of both shapes were pseudo-randomized within each episode such that:

- Each shape (square or diamond) was the most frequently rewarded shape as often as the other one;

- The left-positioned shape was the most frequently rewarded shape as often as the right-positioned one.

Reward distributions

Reward distributions (Figure 7.3) determined, for each shape, the proportion of proposed rewards (2, 4, 6, 8, 10 Euros) displayed before each choice. For all three conditions, reward distributions were pseudo-randomized within episode. Because episodes had finite

small episodes lengths, approximations in reward distributions were made when necessary and fully counterbalanced across episodes. All first trials following a reversal were controlled for respecting the reward distribution too. Rewards were drawn separately for each shape so that the two proposed rewards associated with each shape were statistically independent.

Sessions

Subjects underwent the three experimental conditions in three fMRI sessions, with execution order of conditions counterbalanced across subjects (6 possible permutations, Table 7.1). Male/female participants were counterbalanced with episodes sequence, as well as with execution order of conditions (Table 7.1).

Condition/Sequence	First fMRI session	Second fMRI session	Third fMRI session
Participant 1	1A	2B	3C
Participant 2	1A	3B	2C
Participant 3	2A	1B	3C
Participant 4	2A	3B	1C
Participant 5	3A	1B	2C
Participant 6	3A	2B	1C
Participant 7	1C	2A	3D
Participant 8	1C	3A	2D
Participant 9	2C	1A	3D
Participant 10	2C	3A	1D
Participant 11	3C	2A	2D
Participant 12	3C	1A	1D
Participant 13	1B	2D	3A
Participant 14	1B	3D	2A
Participant 15	2B	1D	3A
Participant 16	2B	3D	1A
Participant 17	3B	1D	2A
Participant 18	3B	2D	1A
Participant 19	1D	2C	3B
Participant 20	1D	3C	2B
Participant 21	2D	1C	3B
Participant 22	2D	3C	1B
Participant 23	3D	1C	2B
Participant 24	3D	2C	1B

TABLE 7.1: Participants were counterbalanced for execution order of conditions (1,2,3) and episodes lengths sequence (A,B,C,D).

Therefore all subjects were homogeneously represented regarding each condition and each episodes sequence.

Breaks

Each session consisted of four runs, interrupted by three breaks. Break duration was up to the subject. The break pseudo-randomly happened two, four or six trials before the end of the run's last episode. Breaks positions were counterbalanced across sessions and across subjects, independently of the other constraints. Subjects were instructed that there was no particular meaning to the moment of break apparition; such that they could not infer any rule or pattern related to the breaks. At each break, the average gain per trial was displayed as well as a "score to beat", for motivational purposes. The score to beat was the performance of a Bayesian model slightly degraded, but the subject was actually told that it was the average score of the best participants so far.

7.1.3 Trial Structure and Jittering

Each trial lasted on average 8100 ms (Table 7.2). Stimuli were displayed for 2500 ms during which subjects could respond at any moment, pressing one of two response buttons within the machine. The stimuli presentation duration, 2500 ms, was calibrated based on the two pilot studies. If participants did not respond within 2500 ms, the trial was lost. Jitters were introduced to temporally decorrelate the various events: stimuli presentation, choice, feedback. As soon as a response was provided, the chosen option remained on the screen until and during feedback was provided. After a first jitter (mean 2.1 s, range 0.1-4.1 s), feedback was provided during 1000 ms: the value of the obtained reward was displayed in the screen centre if the trial was gained, or 0 Euros was displayed otherwise. A black screen separated each of the trials (second jitter, mean 2.5 s, range 0.5-4.5 s). Chosen shape, obtained reward and reaction time were recorded for each trial.

STIMULUS	Jitter ISI	FEEDBACK	Jitter ITI
2.5 s	2.1 s on average	1 s	2.5 s on average

TABLE 7.2: Trial structure and timing.

Jitters were pseudo-randomly drawn from a uniform distribution such that within a run average jitter was close to mean value 2.1 or 2.5 s. They were also controlled for all first and second trials following reversal, assuring that no timing bias emerge around reversals. However, jitters were controlled within run but not controlled within episode.

7.1.4 Participants

25 healthy individuals (13f/12m, aged 20–25 years) were recruited in Paris, France through an internet database (http://expesciences.risc.cnrs.fr/). They all gave written informed consent (approved by the French INSERM Ethics Committee) throughout a short medical interview, during which the following conditions of were assessed by a MD:

- No neurological or psychiatric history (head trauma, epilepsy)
- No attentional or memory disorder
- No dyslexia
- No chronic disease
- No medication (except contraception)
- Normal or corrected-to-normal vision, no fatigue in front of a screen or ophthalmic migraines
- No drug consumption, in particular during the week preceding the experiment, no alcohol consumption 24 hours before the experiment
- Being right-handed with right-handed parents and right-handed siblings (to get rid of lateralized motor activations in fMRI data)
- No tattoo, no metallic objects in the body (piercing, hearing aid, dental cavity filling, brace, pacemaker, prosthesis, screws)
- No students in psychology, neuroscience or cognitive science (being possibly biased).

They participated in three fMRI sessions that took place on three separate days. Each session was separated from another from one day to eight days. Subjects received written minimal instructions about the task and were instructed that payoffs could vary according to their own gains during the task. More precisely, they were told they would receive a minimal amount plus a bonus calculated on their performance during 2% of the trials randomly selected. This manipulation ensured that they would treat each single trial with equal involvement, and hopefully maintain a high enough motivation along the whole session. However in the end, all participants received 240 Euros for their participation in the three sessions.

7.1.5 Training

After receiving written instructions (original document provided in Appendix B, in French) and before entering the MRI, subjects performed a short training on the task, on the specific condition of the session. The training design was the same as in the real experiment. It included 50 trials, which corresponded to three episodes/two reversals.

7.1.6 Debriefing

An informal debriefing took place after the last session (original document provided in Appendix C). After possible comments from their own initiative, subjects were asked, among other questions, 1) whether they noticed any differences between the three conditions, and 2) whether they tended to make their decisions rather according to values or to shapes (stimuli).

7.2 Behavioral Analyses

All statistical analyses were performed under MATLAB R2011a and SPM8 (Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, UK).

7.2.1 Learning Curves

Two learning curves (Figure 8.1) were studied:

- Choice proportion of the most frequently rewarded shape, averaged over subjects and over episodes, and plotted against trial number after a reversal
- Choice proportion of shape with highest expected value, averaged over subjects and over episodes, and plotted against trial number after a reversal. In this kind of economic task, expected value is defined as the probability of obtaining the proposed reward x magnitude of the proposed reward. If subjects were optimal, they would choose, at each trial, the shape with the highest expected value.

The first episode of each session was removed, because it did not consist of an actual reversal. Corresponding fMRI data will be removed too.

We also examined the distribution of choices given the proposed rewards before choice, independently of which shape was the most frequently rewarded one (only given the proposed rewards). The results are provided in Figure 8.2.

At this point, three subjects were removed from the original group for performing at chance level, with stereotypical behaviors; for example, choosing systematically the highest of the two proposed rewards, totally ignoring the shapes. Figure 7.4 showed that these three subjects did not improve their performance over the course of an episode.

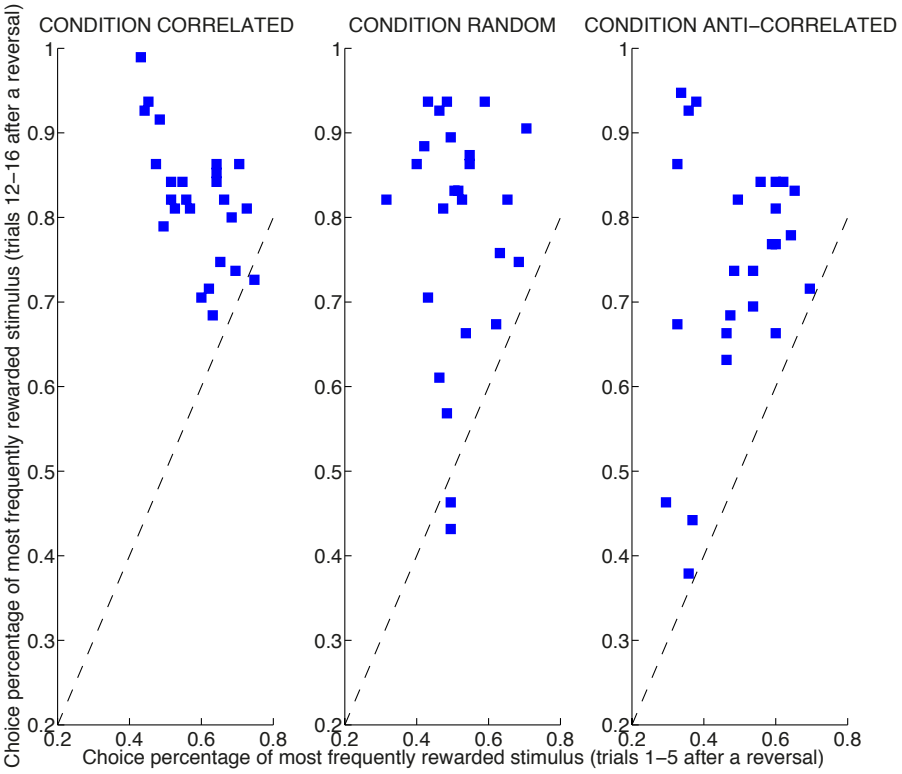


FIGURE 7.4: Subjects falling below the diagonal did not improve their performance over the course of a behavioral episode.

Finally, reaction times were plotted against trial after a reversal and compared across conditions.

7.2.2 Logistic Regressions

We performed behavioral logistic regressions over the choices sequence (dependent variable), in order to scrutinize what variables contributed to choice. Possible explanatory variables were:

- Theoretical probability of being rewarded for each shape (0.8/0.2)
- Proposed rewards displayed before choice

- Subsequent expected values associated with each shape (probability of obtaining the proposed reward multiplied by proposed reward)
- Reward received at previous trial (parametric)
- Binary feedback (rewarded/not rewarded) at previous trial
- Chosen and unchosen proposed rewards at previous trial (in relation to possible counterfactual thinking and regret)

Logistic regressions were performed in full variance analysis, meaning that common variance between regressors will be washed out. The observed remaining contribution of each regressor was thus specific to each regressor. In other words, it means that changing the regressors order did not change the results. All regressors were z-scored before entering the regression, meaning that the relative contribution of each can be compared. Two subjects were removed from the logistic regressions for being group outliers i.e. their only significant contributive regressor was the previous trial outcome.

7.3 Computational Modeling

To establish the functional and computational foundations of the dissociation between affective and informational values, we examined and developed mathematical models of learning and decision. The aim of such cognitive modeling is to understand the mechanistic computations underlying behavior, i.e. hidden variables that are not directly visible in behavioral data. We emphasize the importance of testing various realistic alternative models. The model accuracy in explaining subjects' behavior is always relative. The selected model is best only among a limited number of candidate models, which are never exhaustive.

The generative model of the task in Figure 7.5 corresponds an optimal Bayesian model and is formally discussed in Appendix D. The following sections describe both a Bayesian inference model, corresponding to the generative model of the task (Figure 7.5), and a potential Reinforcement Learning model.

7.3.1 First class of models

No treatment of the informational value of rewards presented before choice

Reinforcement learning model. According reinforcement learning algorithms (Sutton and Barto, 1998 [108]), a "Q value" i.e. average expected value for each option was

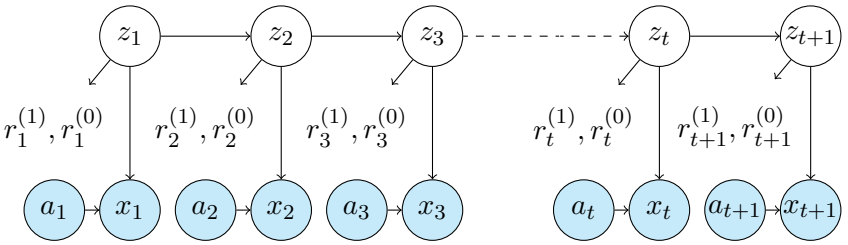


FIGURE 7.5: Generative model of the task: z_t: underlying hidden state; $r_t^{(1)}, r_t^{(0)}$: proposed rewards before choice; a_t: action performed; x_t: feedback observed.

maintained. More precisely, the chosen Q value was updated at each trial according to a prediction error modulated by a learning rate α, in accordance with the Rescorla-Wagner rule [109]. The unchosen Q value was not updated:

$$Q_{t+1}^{(j)} = \begin{cases} Q_t^{(j)} + \alpha(x_t r_t^{(j)} - Q_t^{(j)}) & \text{if } a_t = j, \\ Q_t^{(j)} & \text{otherwise,} \end{cases} \tag{7.1}$$

for $j \in \{0, 1\}$, with learning rate α, with $x_t = 1$ if the subject was rewarded with r_t, and $x_t = 0$ otherwise.

We implemented a slightly modified version of the classic Rescorla rule to allow fair comparison of this standard RL model with more sophisticated models. We did not simply implement action selection on the current expected returns $Q_{t+1}^{(j)}$. Indeed, in our protocol, the potential rewards to gain were displayed before each choice. Therefore, to take into account the influence of the proposed rewards displayed before each choice, we biased the current expected returns by:

$$\tilde{Q}_t^{(j)} = w r_t^{(j)} + (1 - w) Q_t^{(j)} \tag{7.2}$$

then perform action selection on the basis of these biased expected returns. The bias corresponded to a "what if" step that is only performed for action selection but not taken into account for inference.

Hence, this model included two free parameters: the learning rate α and the bias w over current Q values, as described above. At the beginning of each session, Q values were initialized at their mean value (6 Euros). Fitting the initial Q values as a free parameter did not significantly improve the model. Importantly, this RL model consisted in a continuous trial-by-trial update; it assumed no task structure, specifically, no structure related to the reversals.

Normalized Reinforcement learning model. Building on the classic RL model above, the normalized RL model updates both chosen and unchosen Q values at each trial. More precisely, the normalized RL assumed that:

- If no reward was given, a reward would have been given for choosing the other option

- If a reward were given, choosing the other option would have led to no reward.

Implicitly, this model assumed an underlying structure in the task, which was that the two options would be opposite. But it did not assume either any structure related to reversals.

Bayesian model with no processing of the informational value – This model was a particular sub-case of the Bayesian model described in the next section, in which one step was omitted (the belief update by the informational values conveyed by proposed rewards before decision). This model was necessarily sub-optimal because it was blind to the three experimental conditions, which corresponded to three different reward distributions.

7.3.2 Second class of models

Treatment of the informational value of rewards presented before choice

The following models were based on the classic economic assumption of utility maximization (Kahneman andTversky, 1984, 1979 [170]; [98]). In our task, optimal choice was to choose the option with the largest expected utility. Computing an expected utility corresponded to the optimal combination of information about probabilities and about rewards. This class of models theoretically adapted faster when a reversal occurred, thanks to the use of the informational values carried by proposed rewards before choice.

Bayesian model. More precisely, the Bayesian model monitored a *belief* about how shapes mapped onto outcome contingencies (i.e. in this task, which shape was the most frequently rewarded one). A key feature of this second class of models was that informational value from the proposed rewards presented before choice was extracted and used to update the belief before choice (as a likelihood in a Bayesian framework). An expected value was then computed for each shape: belief multiplied by proposed reward. After choice, a binary feedback (win/lose) was extracted from the received outcome and used to update the belief using Bayes rule. A volatility parameter ν determined the reversals frequency between shapes, meaning that the hidden state underlying shapes changed with probability ν. Formal description and inferences steps of this Bayesian model are provided in Appendix D.

Hence, this Bayesian model included the following free parameters:

- Volatility ν; since episodes lengths were pseudo-randomized, volatility was assumed to be constant across the experiment.

- Probability q of obtaining a reward for having chosen the most frequently rewarded shape. In our experimental design, q was set equal to 0.8. In other words, once an option was chosen, feedback was provided probabilistically (q or *1-q*) given the chosen option.

- Slope γ of the exponential distribution used to generate proposed rewards at each trial, separately for each of the three experimental conditions. Values from the experimental design were γ *random* = 0 for condition random, γ *correlated* = 0.13 for condition correlated and γ *anti-correlated* = - 0.13 for condition anti-correlated. In our case, $\gamma = 0$ indicated that the proposed rewards were uninformative about the hidden state (condition random), $\gamma > 0$ implied that choosing higher rewards was overall more profitable (condition correlated), and $\gamma < 0$ implied that choosing lower rewards was overall more profitable (condition anti-correlated). These generative parameters were chosen based on the two pilot studies of the task, so that the distributions were neither too difficult nor too obvious.

At the beginning of each session, beliefs were initialized at their mean value (0.5): subjects had no reason to prefer one of the two shapes at the beginning. We checked that fitting the initial belief as a free parameter did not significantly improve the model.

Distortions model. Prospect theory (Kahneman and Tversky, 1984 [170]) states that subjects' choices are based on maximizing an expected value (= expected utility) and that subjects' deviations from rationality can be explained by distortions in their internal probabilities and rewards representations (Tversky and Kahneman,1974 [171]; Kahneman and Tversky, 1979 [98]). In our task, these two dimensions were at stake. Specifically, each shape was associated with: (1) a certain probability of leading to a reward; (2) a proposed reward to be potentially gained. Therefore, we built on the above Bayesian model, modifying only the beliefs and the proposed rewards by distortions. All possible types of distortions were examined (concave, convex, sigmoid, inverse sigmoid) as Zhang and Maloney formalized them [172], separately for beliefs and for proposed rewards. The addition of distortions resulted in four more free parameters: slope and fixed point for probabilities, slope and fixed point for rewards.

7.3.3 Third class of models

Treatment of the informational value of rewards presented before choice but no expected value computation

Mixed model: beliefs system and affective value systems. In this model, choice was made over a mixture of beliefs (Bayesian inference system) and affective values (RL system), but not in the form of an expected value computation. Instead of computing an explicit expected value, choice was based on a combination of two systems:

- A reinforcement learning system, processing the rewards affective value, as described in our first class of models;

- A Bayesian inference system, processing the proposed rewards informational value, by building beliefs about states (which shape was the 'correct' one, i.e. the most likely to lead to obtain a reward), as described in our second class of models.

According to this model, a belief about how shapes map onto outcome distributions and an affective value were combined to make a decision (Figure 7.6). More precisely, the belief was a prior belief from the past. When subjects observed the proposed rewards before choice, their prior belief was revised by the information value conveyed by proposed rewards (likelihood in a Bayesian framework). In parallel, the affective value was the sum of reinforcement value from previous trials and proposed rewards at decision time. Subjects made a decision by combining these two systems, and subsequently observed an outcome. Given this outcome, the beliefs were updated by Bayes rule, while the affective values were updated by standard reinforcement learning (Figure 7.6).

The mixed model contained the free parameters of a RL model and the free parameters of a Bayesian model, as described above. Additionally, it encompassed a weight parameter ω arbitrating between the belief system (Bayesian) and affective values system (RL). Given the high-dimensional parameters landscape to adjust to data, we ensured our fitting procedure was reliable (see section Fitting procedure below).

Mixed model: expected values (Bayesian) and affective values (RL). Importantly, we checked that a mixed model combining (1) an expected value (belief multiplied by proposed rewards) from a Bayesian system and (2) a reinforcement values from a RL system consistently fitted less well the behavior than the previous mixed model.

7.3.4 Action selection

A general strategy for action selection was to stochastically select an action a_t according to the softmax policy (Luce, 1977 [104]), with parameters β and ϵ:

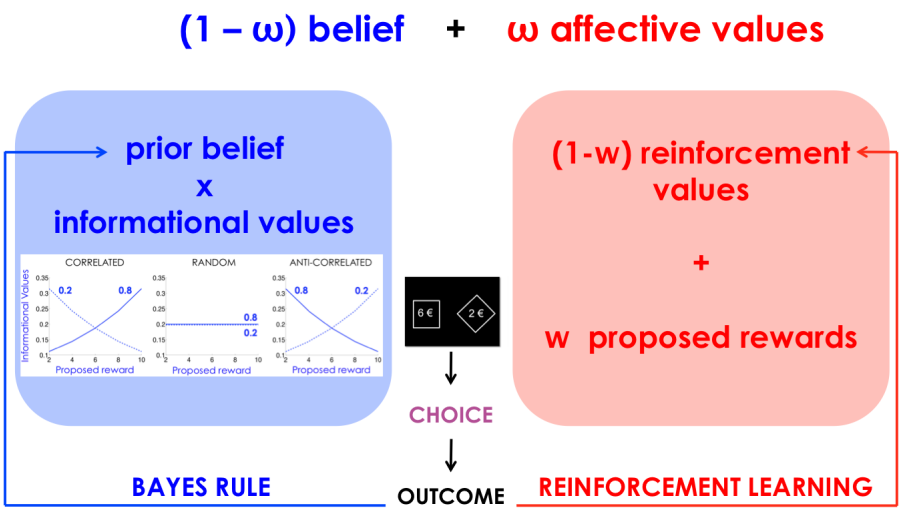


FIGURE 7.6: Schematic representation of the computations performed by the mixed model combining a beliefs system and an affective values system. Details are provided in the main text.

$$p(a_t = 1) = \frac{\epsilon}{2} + (1 - \epsilon)\frac{e^{\beta Val1_t}}{e^{\beta Val1_t} + e^{\beta Val2_t}}, \tag{7.3}$$

with $Val1_t$ and $Val2_t$ being the expected values for choosing option 1 and option 2 respectively, and β the softmax inverse temperature, allowing for exploration towards the lower-valued action. The optimal strategy is obtained when β tends towards infinity. In that case, the subject would pick, at each trial, the option with the largest expected value. The term with ϵ models the lapses proportion, with $\frac{1}{2}$ being the random choice probability. After fitting procedure (see below), we obtained large values of β and low values of ϵ for the mixed model, meaning that the mixed model was a reliable predictor of subjects' actions.

7.3.5 Fitting procedure

The fitting procedure objective is to find the set of free parameters that best fits each subject's data. The number and nature of parameters depended on each particular model. For adjusting the parameters to the behavioral data (here, subject' choices), we maximized the model log-likelihood (LLH):

$$LLH = \prod_t \log(p_t)$$

Where p_t is the probability that the model would have chosen the same action as the subject at trial t.

Log-likelihood was relevant in our case as it is a sensitive measure, as compared to least-squares minimization for example, and also because it equally takes into account all actions.

Parameters were individually adjusted, because we assumed different subjects could have different learning rates, different distortions... Although it has been shown that in certain cases parameters could vary across time (Khamassi et al., 2013 [173]), reflecting online adjustments, here we did not aim at describing the precise learning dynamics. Therefore, we made the approximation of constant parameters during a session. Even if certain parameters could evolve through time, our parameters were supposed to represent individual features, which might be related to each subject's particular neurophysiology. The parameters number and nature of the best-fitting model are provided in Table 8.1. For each subject, all three conditions were fitted as a whole, but certain parameters varied across the three conditions (reward distribution slope γ in Bayesian model) whereas all other parameters remained constant across the three conditions (e.g. volatility). Only one weight parameter was used to fit the three conditions. Having three different weight parameters did not significantly improved the model (see Results).

Various methodologies were considered to find the set of parameters that best fitted the data. For models with few free parameters, all procedures were generally able to converge towards the best set of parameters. By contrast, for high-dimensional parameters spaces, the problem was more complex.

Grid search. A first possibility was to explore a grid by taking a finite number of discrete values for each parameter. Then, for each parameters combination, the log-likelihood was calculated and the maximum log-likelihood was retained. This method presented two main drawbacks. On the one hand, it allowed exploration of only discrete parameter values, whereas all of them could take continuous values. On the other hand, as the number of parameters increased, the calculation time exponentially increased (the curse of dimensionality).

Gradient ascent. Another possibility was to use a gradient ascent. The MATLAB Optimization Toolbox contains several tools to find the maximum of constrained non-linear multivariable functions, by calculating partial derivatives of the function. Here the function was our model's log-likelihood, with multiple parameters. For any starting point, this mathematical procedure was able to converge to the closest, local maximum log-likelihood and provided the best set of parameters associated with this likelihood. In that case, the exploration was continuous, but the main problem was that the algorithm could be stuck in a local maximum and not find the global maximum, even when tuning the algorithm's options.

Slice sampling. Eventually, we used a slice sampling procedure. This method has a high computational cost but it presents advantages for high-dimensional parameters space (Bishop, 2006 [107], chapter 11.4). It allows for checking a posteriori the variance and the shape of each parameter posterior distribution. This constitutes a way of evaluating the parameters estimation accuracy, and ensuring that the whole parameters landscape was explored. More precisely, 100,000 samples were drawn and an additional gradient ascent was performed on the best sample (Optimization Toolbox, MATLAB). Using three different starting points drawn from uniform distributions for each parameter, and drawing 200,000 samples for each starting point, provided the same fit quality. A posteriori, cross-correlation diagrams were drawn for each parameter and for each model, to check that the samples were independent enough from each other. Generally, it is assumed that 10 times n independent samples are necessary to obtain an accurate parameter's average estimation. For illustrative purposes, Figure 7.7 shows an example of such diagram for all parameters, for an example subject. Each line corresponds to a parameter. We can see that starting from n = 3,000 here, samples were independent enough. In that case, 30,000 samples would thus be sufficient for parameters estimation to be accurate. However, we kept the same number of samples (100,000) for all subjects and all models for consistency. The fact that the samples were independent enough ensures that the parameters average estimate was reliable.

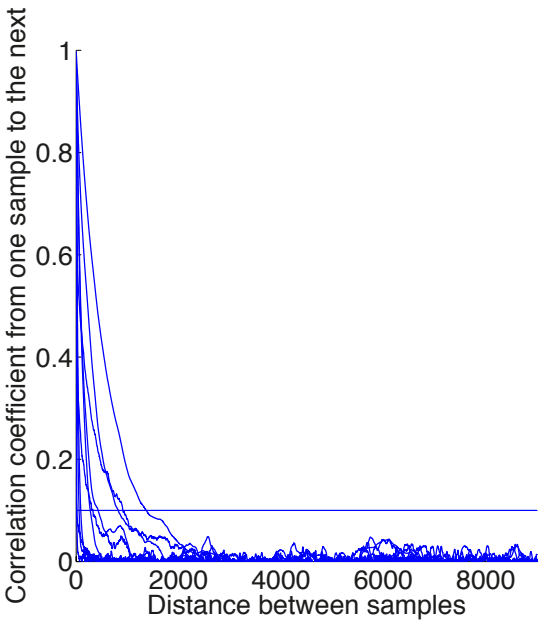


FIGURE 7.7: Cross-correlation diagram.

Additionally, parameters were plotted by pairs, grouping all samples (cloud plots in Figure 7.8). The closer together the dots are in the cloud, the narrower the posterior

variance was for each parameter. An illustrative example of a relatively good estimation for an example subject is displayed in the left panel, whereas an example of two parameters whose posterior distributions were partly linked is displayed in the right panel (Figure 7.8).

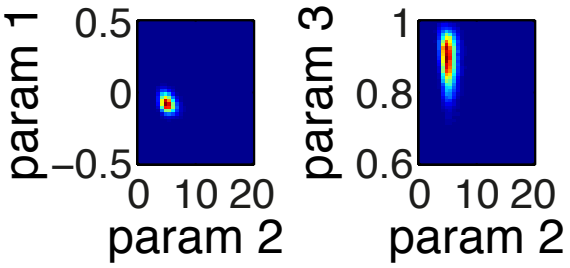


FIGURE 7.8: Pairs of parameters' samples. The closer together the clouds of dots are, the narrower the posterior variance for each parameter is.

Another confirmation is the log-likelihood evolution across all samples. Figure 7.9 displays three examples of results for an example subject. An example of good convergence is illustrated in the left panel. After a burn-in phase (removed), the log-likelihood stabilized close to its maximum value, with oscillations around this maximum (unimodal distribution). An example of a bimodal distribution is illustrated in the middle panel. It means that the log-likelihood oscillated between two modes with no clear unique maximum. Finally, an example of non-convergence is illustrated in the right panel. It means that there was probably another mode that was not reached yet.

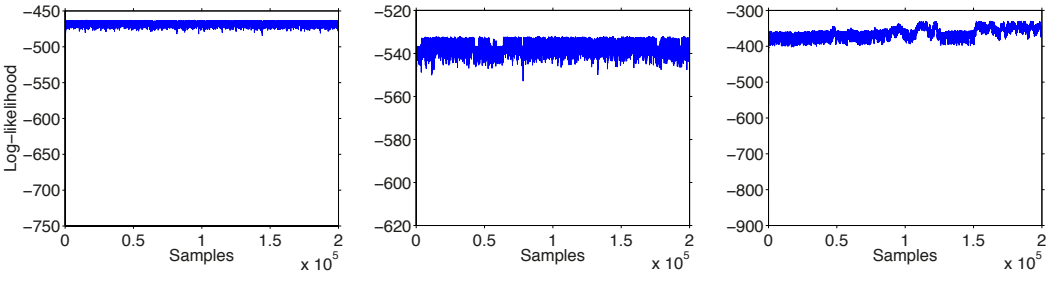


FIGURE 7.9: Log-likelihood evolution through all samples during slice-sampling procedure. Left panel: Example of good convergence. Middle panel: Example of bimodal distribution. Right panel: Example of non-convergence.

With that procedure, we were thus able to identify, for each model, the free parameters set that best fitted the subjects' behavioral data. The next step was to select the best model among several alternative models, in terms of both fit goodness and parsimony.

7.3.6 Model selection

A crucial point in modeling is model comparison. Indeed, we could imagine a model that has a very high likelihood but that is not capturing well what subjects are doing. By contrast, a model with a large number of parameters could capture very well what subjects are doing, but is actually over-fitting the data (Hawkins et al., 2004 [168]). To prevent from this possibility, we evaluated both qualitative and quantitative measures for each model. We emphasize the importance of presenting models simulations to give an idea of the model qualitative behavior and to support its relevance for explaining the participants data.

7.3.6.1 Quantitative measures

The log-likelihood obtained for each fit gave an index of how well the model predicts the subject's choices. However, for a given model, the higher the number of free parameters added, the higher the log-likelihood, but it can be an artificial increase. To take into account the model complexity, we used the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC).

$$BIC = -\,2\ LLH\ +\ k\ \ln(n)$$

$$AIC = 2\ k - 2\ LLH$$

with k the number of the model free parameters. The BIC penalizes more the extra parameters because it accounts for the number of observations n (= number of trials) used to fit the data. Quantitative measures can be misleading though, as log-likelihood overweighs low probability actions into the global calculation. Therefore, the following measures were also critical to examine whether the model qualitatively predicted the subject's choices.

7.3.6.2 Qualitative measures

Crucially, we assessed whether our models qualitatively reproduced subjects's behavior. To that aim, model's simulations were performed. Taking the fitted parameters of each subject, the model was run as if it was a subject. It was then possible to study its choices sequence similarly as for participants (cf. Behavioral Analyses). In particular, two behavioral measures were examined. First, we reproduced for each model's simulation the learning curves computed in Figure 8.1:

- Choice proportion of the most frequently rewarded shape
- Choice proportion of shape with highest expected value

Moreover, the Figure 8.2 showing the choice proportion of each proposed reward was reproduced for model's simulations, again to assess whether model's choices were consistent with subjects' choices.

7.4 Neuroimaging

We investigated whether the two systems (beliefs and affective values) in our best-fitting computational model had distinct neural bases using functional MRI.

7.4.1 fMRI acquisition

fMRI volumes were acquired on a 3T Siemens Trio at the Centre de Neuroimagerie de Recherche (CENIR) within the hospital La Pitié Salpêtrière in Paris. Acquisition parameters were TR = 2 s, TE = 25 ms, 431 repetitions per run, 4 runs of 14'28 based on the longest run duration. 39 slices of 2 mm thickness were acquired by sequential-descending order, flip angle 75°, with slice number 39 as reference slice. Before the first trial, 2 TR of baseline recording were acquired, to allow for slice-timing correction (see Pre-processing below). EPI were 30° tilted to minimize signal drop around the orbitofrontal cortex (Deichmann et al., 2003 [174]). For acquisition voxel size was 2.5 x 2.5 x 2.5 mm^3. The experiment was projected on a mirror settled on a 32-channels head coil. Subjects provided their responses through two MRI-compatible response buttons, one in each hand. In addition, T1 anatomical images as well as FieldMaps were acquired. Diffusion Tensor Images were also recorded for the purpose of later comparing classic DTI sequences vs. DTI multiband sequences; nevertheless, no anatomical connectivity analysis is provided in this thesis.

7.4.2 fMRI pre-processing

Data were preprocessed using SPM8 (Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, UK). Images were first reoriented such as the origin corresponds to the anterior commissure. Inverse coordinates (up, right, forward) were recorded. Below are listed the procedures in the order in which they were performed.

Slice-timing correction (temporal correction) – The purpose of this correction was to have a Gaussian distribution of noise in the data. Between two volumes, an interline

acquisition was added and then it was considered that all were acquired at the same time. This correction can be particularly critical to perform Dynamic Causal Modeling.

Realignment and motion correction (spatial correction) – All displacements superior to 1 mm were corrected, taking the first image as a reference point, by doing a rigid transformation containing three translations (x, y, z) and three rotations (pitch, roll, yaw). Six movement regressors were thus extracted and will be used as non-interest regressors in the GLM. These movement variations have to remain low enough; movements up to 3-5 mm or 3-5° were accepted, otherwise images were checked by hand.

Segmentation and spatial normalization – The anatomical image T1 was normalized into white matter, grey matter and cerebrospinal fluid. The same normalization applied to the T1 was applied to all functional images.

Co-registration – This step aims at linking the functional images (T2) to the anatomical image (T1) in the same space. The mutual information diagram should be less scattered after this correction (illustrated in Figure 7.10).

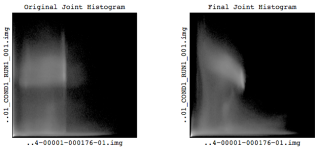


FIGURE 7.10: Preprocessing: coregistration. The mutual information diagram was less scattered after this step.

Spatial normalization – The purpose of this non-rigid deformation is to link functional images to a template brain. This was the way we are able to compare activations between subjects, despite inter-individual brain morphology variability. We applied all the calculated deformations to take back all the functional series to a template. The normalized data voxel size was 3 x 3 x 3 mm.

Smoothing – This correction is supposedly the most efficient one. It diminishes the noise by averaging the signal in each voxel by a Gaussian kernel according to the signal intensity in adjacent voxels. Then the average signals look more alike: a strong effect in one voxel will be less intense after smoothing but its spatial extension will increase. This refinement is reasonable since the activities of two adjacent voxels are very correlated, and two adjacent regions might have a functional similarity. Besides, the vascular system is quite "blurred". Finally, the spatial smoothing extent has to be carefully chosen since a too large filter could result in activations outside the brain! Here we chose Full Width at Half Maximum = 6 mm (Gaussian kernel width).

At this point, one additional subject was removed from the group because of excessive head movement (up to 25 mm in translation and 5° in rotation). Despite correction some frontal cortex voxels were missing. Consequently, all following fMRI maps were obtained for 21 subjects.

7.4.3 fMRI: Model-based approach

A presentation of the model-based approach is illustrated in Figure 7.11. The computational model was our access to brain mechanisms not directly visible in the behavior. Using fMRI allowed us to probe the biological implementation of our best-fitting computational model.

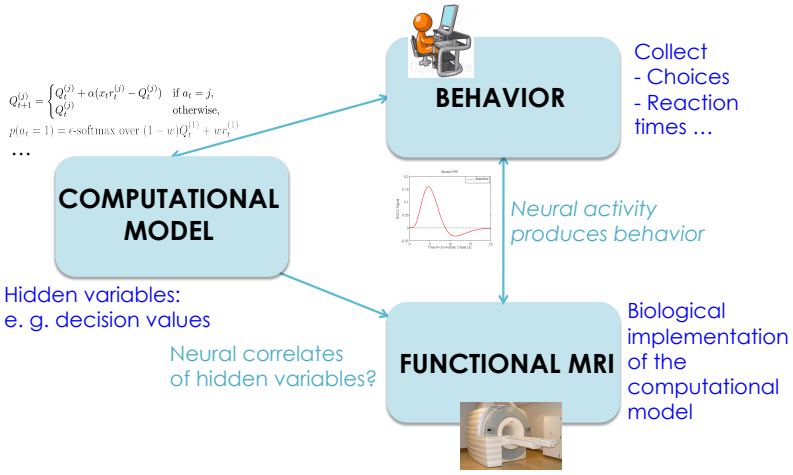


FIGURE 7.11: A schematic illustration of the model-based approach.

The model-based approach interest is to provide a mechanistic explanation (computations, representations) of the behavior, not only an information of localization within the brain. Indeed, it can probe not only whether experimental conditions vary but also how and why they vary (Mars et al., 2012). The model-based fMRI can answer difficult questions for the model-free approach, such as the task information neural encoding, or how specific brain signals modulate model parameters.

7.4.4 fMRI: General Linear Model

A General Linear Model (GLM) was conducted to analyze the fMRI data. A first-level analysis was performed using model-based fMRI (O'Doherty, 2007 [175]). Essentially, this method allows to identify regions that specifically correlate with a model's variable. Rather than solely identifying locations, model-based fMRI informs about the cerebral

implementation of the cognitive mechanisms that our best-fitting computational model (mixed model) described. In short, this model described decision-making as a mixture of two systems (Figure 8.18):

- A belief system, that processed probabilities of obtaining a reward
- A reinforcement learning system, that processed affective values

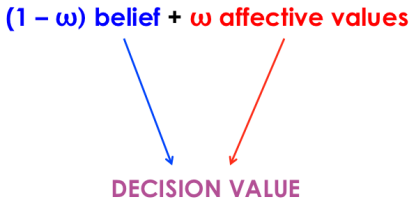


FIGURE 7.12: Best-fitting mixed model included both a belief system and an affective value system that were combined to make a decision.

First, we examined the neural correlates of decision values i.e. mixture of beliefs and affective values (GLM1). Critically, we then addressed the question whether these two systems had distinct neural bases (GLM2). Further dissociations within each system were investigated in GLM3.

More precisely, two main regressors of interest were incorporated: **STIMULUS** and **FEEDBACK**, each including several parametric modulations that are detailed below. All three conditions were pooled into the GLMs, since the differences between conditions were captured by the computational model. Stimulus and feedback events at each trial were modeled as a Dirac (Event-related design). Modeling events as a 2-seconds block produced similar parametric maps.

Crucially, a full variance analysis was performed (sometimes named "unique variance analysis"). Specifically, we deactivated the default SPM option that orthogonalized the parametric modulators in the order they appear. Consequently, all common variance between modulators was placed in the residuals. Therefore, the observed activations were specific to each parametric modulator. However, certain parametric modulators were manually orthogonalized (Gram-Schmidt orthogonalization) before they were entered into SPM (details below). All parametric modulators were z-scored before they were entered into SPM. Lapses (absence of response) were modeled in two separate regressors (lapses at onset stimulus, subsequent absence of feedback at onset feedback). Across subjects, lapses consisted on average of 0.6% of all trials. Regressors of no interest included six movement parameters from the realignment procedure, as well as a regressor modeling each run.

7.4.4.1 GLM1: Decision Values

The following parametric modulations corresponding to variables from our best-fitting computational model were included:

STIMULUS - onset of stimuli apparition. Given fMRI temporal resolution, we cannot be sure that this onset precisely captured decision time, but it covered at least part of the decision time window.

- Parametric modulation 1: Decision value chosen – decision value unchosen
- Parametric modulation 2: (Decision value chosen – decision value unchosen)^2, orthogonalized with Parametric modulation 1

FEEDBACK - onset of feedback apparition. Parametric modulations corresponding to variables from our computational model were included:

- Parametric modulation 1: Belief chosen (before feedback reception)
- Parametric modulation 2: Reward received, associated with chosen shape (which could be either chosen reward or 0 euros)
- Parametric modulation 3: Reward chosen (proposed reward that was chosen, before outcome is revealed), orthogonalized with Parametric modulation 2
- Parametric modulation 4: Binary feedback (x_t in the model; coding win vs. lose), orthogonalized with Parametric modulation 2 and 3

In particular, at stimulus onset, we regressed a linear effect (Decision value chosen - decision value unchosen) and a quadratic effect ((Decision value chosen - decision value unchosen)^2, orthogonalized on the linear effect), under the following interpretative hypotheses.

Positive Linear Effects. A brain region showing a positive linear effect as in Figure 7.13 corresponds a region that is more activated when relative chosen value is higher. This corresponds to the pattern of a region encoding expectations associated with chosen shape (action outcome expectation).

Negative Linear Effects. A brain region showing a negative linear effect as in Figure 7.14 is a region in which activity decreases when relative chosen value increases, which might reflect the evidence accumulation process for decision or the unchosen option value encoding.

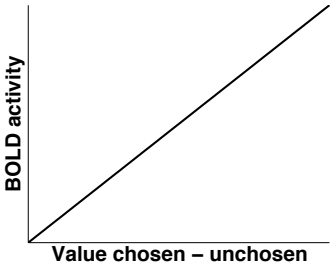


FIGURE 7.13: Pattern of a region showing a positive linear effect, which was interpreted as encoding expectations associated with chosen shape.

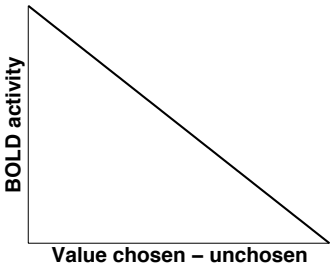


FIGURE 7.14: Pattern of a region showing a negative linear effect, which was interpreted as reflecting the evidence accumulation process for decision.

Positive Quadratic Effects. A brain region showing a positive quadratic effect as in Figure 7.15 is a region that is more activated when chosen and unchosen values are far from each other. This parametric modulator codes for the difference between the two option values, irrespective of choice. This pattern corresponds to a region that encodes pre-choice preferences (unsigned by choice), or a post-choice confidence signal (i.e. the further the values from each other, the higher the confidence in choice).

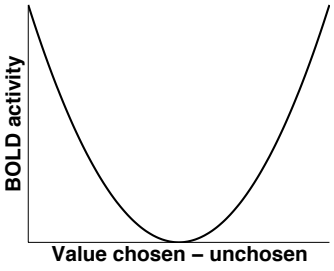


FIGURE 7.15: Pattern of a region showing a positive quadratic effect, which was interpreted as encoding pre-choice preferences.

Negative Quadratic Effects. A brain region showing a negative quadratic effect as in

Figure 7.16 is a region that is more activated when chosen and unchosen values are close to each other. Therefore such a region would be more activated when choice becomes more difficult, corresponding to choice difficulty encoding, uncertainty or would be a region performing action selection.

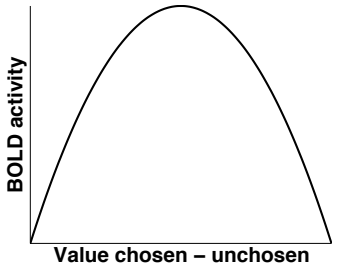


FIGURE 7.16: Pattern of a region showing a negative quadratic effect, which was interpreted as a region performing action selection.

Critically, linear effects were signed by choice i.e. choice-dependent, whereas quadratic effects were unsigned by choice i.e. choice-independent.

7.4.4.2 GLM2: Dissociation belief system/affective values system

The following parametric modulations corresponding to variables from our best-fitting computational model were included:

STIMULUS - onset of stimuli apparition.

- Parametric modulation 1: Belief chosen – Belief unchosen (after update by informational values)
- Parametric modulation 2: (Belief chosen – Belief unchosen)^2 (after update by informational values), orthogonalized with Parametric modulation 1
- Parametric modulation 3: Q chosen – Q unchosen (after bias by proposed rewards presented before choice)
- Parametric modulation 4: (Q chosen – Q unchosen)^2 (after bias by proposed rewards presented before choice), orthogonalized with Parametric modulation 3

FEEDBACK - onset of feedback apparition. The same parametric modulations were included for GLM1, GLM2 and GLM3.

The above interpretative hypotheses for positive and negative linear and quadratic effects still stand for GLM2 and GLM3.

Importantly, we controlled the consistency of brain activations when including additional parametric modulations coding for:

- Parametric modulation: Reaction times
- Parametric modulation: Stay/Switch, according to whether the subject sticked with the same choice or switched choice as compared with the previous trial

We also examined the shape of reaction times as a function of belief chosen, a variable extracted from our best-fitting computational model. To do that, we binned the data using two different methods:

(1) Bins with fixed bounds but variable number of events per bin. We sorted trials with increasing chosen belief and split them into 10 bins of equal size (from 0 to 1 by steps of 0.1). This resulted in a variable number of trials per bin because in learning paradigms as ours, there is a sampling asymmetry given that subjects more often chose the shape with the higher belief.

(2) Bins with variable bounds but fixed number of events per bin. We sorted trials with increasing chosen belief and we split them into about 10 bins of about 100 trials per bin, which resulted in bins with bounds of variable size.

7.4.4.3 GLM3: Further dissociation within each system

The following parametric modulations corresponding to variables from our best-fitting computational model were included:

STIMULUS - onset of stimuli apparition.

- Parametric modulation 1: Prior belief chosen – Prior belief unchosen (before update by informational values)
- Parametric modulation 2: (Prior belief chosen – Prior belief unchosen)^2 (before update by informational values), orthogonalized with Parametric modulation 1
- Parametric modulation 3: Informational Value associated with chosen proposed reward – Informational Value associated with unchosen proposed reward. Informational value is the quantity by which beliefs are updated when proposed rewards to gain are displayed before choice (likelihood in the Bayesian system)

- Parametric modulation 4: (Informational Value associated with chosen proposed reward – Informational Value associated with unchosen proposed reward)^2, orthogonalized with Parametric modulation 3
- Parametric modulation 5: Q chosen – Q unchosen
- Parametric modulation 6: (Q chosen – Q unchosen)^2, orthogonalized with Parametric modulation 5
- Parametric modulation 7: Chosen proposed reward – unchosen proposed reward
- Parametric modulation 8: (Chosen proposed reward – unchosen proposed reward)^2, orthogonalized with Parametric modulation 7

FEEDBACK - onset of feedback apparition. The same parametric modulations were included for GLM1, GLM2 and GLM3.

Notably, in all GLM, replacing the quadratic regressors by absolute values instead of squares led to very similar results.

Importantly, the results consistency was assessed when including additional parametric modulations: reaction times (Grinband et al., 2008 [176]) and stay/switch trials. **Variance inflation factor** (VIF) was calculated for each parametric modulation to ensure collinearity between all parametric modulations was small enough (Fair et al., 2006 [177]).

Accordingly, quadratic effects were unsigned by choice, so they would reflect pre-choice/choice-independent variables. On the contrary, linear effects were signed by choice, so they would rather reflect post-choice/choice-dependent variables.

All parametric models were regressed against BOLD signal. Beforehand, BOLD signal was convolved with the hemodynamic response function to model the activations (intrinsic autocorrelations modeled as autoregressive noise of order 1, high-pass filter with cut-off = 128 s). This first-level analysis was performed for each subject individually. Second-level parametric maps were then obtained for each contrast after a smoothing with a kernel of 8 mm width, for the whole group (21 subjects). Significance threshold was set at $p < 0.005$: this threshold corresponded to a correction for the whole frontal lobe, region with a strong a priori. Despite this uncorrected threshold, it should be highlighted that analysis was performed in full variance, meaning the activations that remained significant were truly selective of each parametric modulator (common variance was placed in the intercept).

7.4.5 Regions of Interest (ROI)

Second-level maps were then computed to identify specific clusters correlating with each parametric modulation. β coefficients were extracted to estimate the correlation strength in a given cluster, taking a sphere of radius 13 mm centered on the ROI peak voxel. Notably, we checked that we obtained very similar statistics whether we averaged either all voxels of the cluster or voxels from a small sphere (diameter = 10 - 20 mm) defined around the cluster peak voxel. Coordinates of the peak voxel in each ROI were given using Montreal Neurological Institute (MNI) atlas. The associated labels were checked using the neuroanatomy data provided in the Duvernoy atlas.

To avoid circularity, we ensured that the ROI definition was made on an independent analysis (Kriegeskorte et al., 2009 [178]). To that aim, a leave-one-out procedure was used to extract β in ROI. More precisely, second-level maps were re-estimated for *n-1* subjects and β were extracted for the last remaining subject in ROI defined from this *n-1* second-level map. The procedure was repeated for each of the n subjects, and β of all subjects were then averaged. Thus, the ROI definition was independent on the statistical analysis made within the ROI. However, we noted that here the obtained statistics did not differ much if we did not use the leave-one-out procedure. The main reason was probably that ROI contained quite similar voxels when defined on 21 vs. 20 subjects.

Another possibility to avoid circularity is thus to use ROI coordinates defined from an independent dataset, using previously published data or meta-analyses (e.g. Sescousse et al., 2013 [179]).

A between-subjects analysis was then performed to examine whether there was a link between the activation strength in a ROI and the value of the mixed model's free parameter ω that weighs the contribution of the belief system in the decision. For each subject, the β in a ROI was plotted against the fitted value of ω and correlation was tested. We report here only the significant correlations; because of the intrinsic noise in fMRI, it is to be noted that it is generally difficult to obtain such correlations when working with human fMRI data.

Comparison of the distribution of beliefs and affective values from trial to trial.

This comparison purpose was to rule out the possibility that affective values varied more rapidly from trial to trial than beliefs (or the converse). Indeed, if it had been the case, the neural dissociation that we observed between choice-independent beliefs and choice-independent affective values could have been related to a rather stable variable

vs. a rapidly changing variable. We extracted, for each subject, the mean of the squared difference between relative chosen belief at t and relative chosen belief at $t-1$, and the mean of the squared difference between relative chosen affective value at t and relative chosen affective value at $t-1$. We then tested whether there was a significant difference between the two (paired t-test). The differences were squared because the direction of the variation between t and $t-1$ did not matter, we were only interested in the amount of variability from trial to trial. We also checked with the median instead of the mean to rule out effects due to outliers.

7.4.6 3D Bins analysis

Univariate analysis showed statistically significant quadratic effects, for both beliefs and affective values, in specific clusters (see Results). The aim of the binning analysis" was to be able to visualize the quadratic effects actual shape in the BOLD signal, in a separate analysis.

This was especially delicate because in learning paradigms such as ours, subjects obviously tended to choose more often the highest of two beliefs/affective values. Consequently, we had an asymmetry; we had fewer trials in which (chosen – unchosen value) was negative (Figure 7.17)). On the positive part of the graph 7.17, linear and quadratic effects will be almost indistinguishable (fMRI being too noisy). By contrast, on the negative part of the graph 7.17, we will be able to distinguish between linear and quadratic effects, but we had fewer data points for this part.

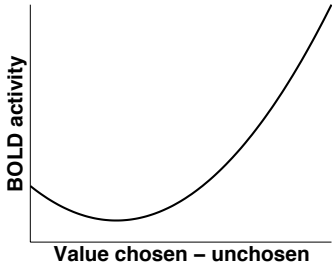


FIGURE 7.17: Schematic example graph with both linear and quadratic effects coexisting. There was an asymmetry between the left and the right part of the graph, due to the fact that, in learning paradigms, subjects typically chose more often the most valued of two options.

In each voxel, BOLD signal was dependent on both beliefs and affective values. Therefore, the trials were sorted and binned according to both beliefs and affective values, on a two-dimensional grid (Figure 7.18). The boundaries of each bin were constant. As a result, we had a various number of events per bin. Beliefs and affective values were

z-scored before sorting and building the bins. The onsets of trials falling in each bin were then collected and a first level analysis was performed to estimate an average brain activity β for each bin and each voxel. From this 3D plot, data was then projected on each dimension, in order to see the signal evolution either according to beliefs or according to affective values. To that aim, we averaged on each dimension, by basically marginalizing over each dimension alone. The mean and the standard error of the mean were calculated across bins. On this raw data, we fitted a degree-2 polynomial. Finally, we de-trended the data according to this polynomial fit in order to observe linear and quadratic terms separately. It means that on Figure 7.18, the sum of the curve "linear" and "quadratic" corresponded to the "overall" effect. The purpose of this bins analysis was to actually visualize the linear and quadratic effects that were detected in the second-level parametric maps.

In this bins analysis, we did not take into account inter-subjects effects. It was a fixed-effects analysis: all subjects were pooled when building the bins. Error bars correspond to the standard error of the mean across bins.

Note: We imposed the same boundaries for the bins for all subjects. Another possibility would have been to bin by quantiles, i.e. impose an identical number of events per bin. Consequently, bins boundaries and bin sizes would have been variable across subjects. This possibility can only be implemented when building one-dimensional bins.

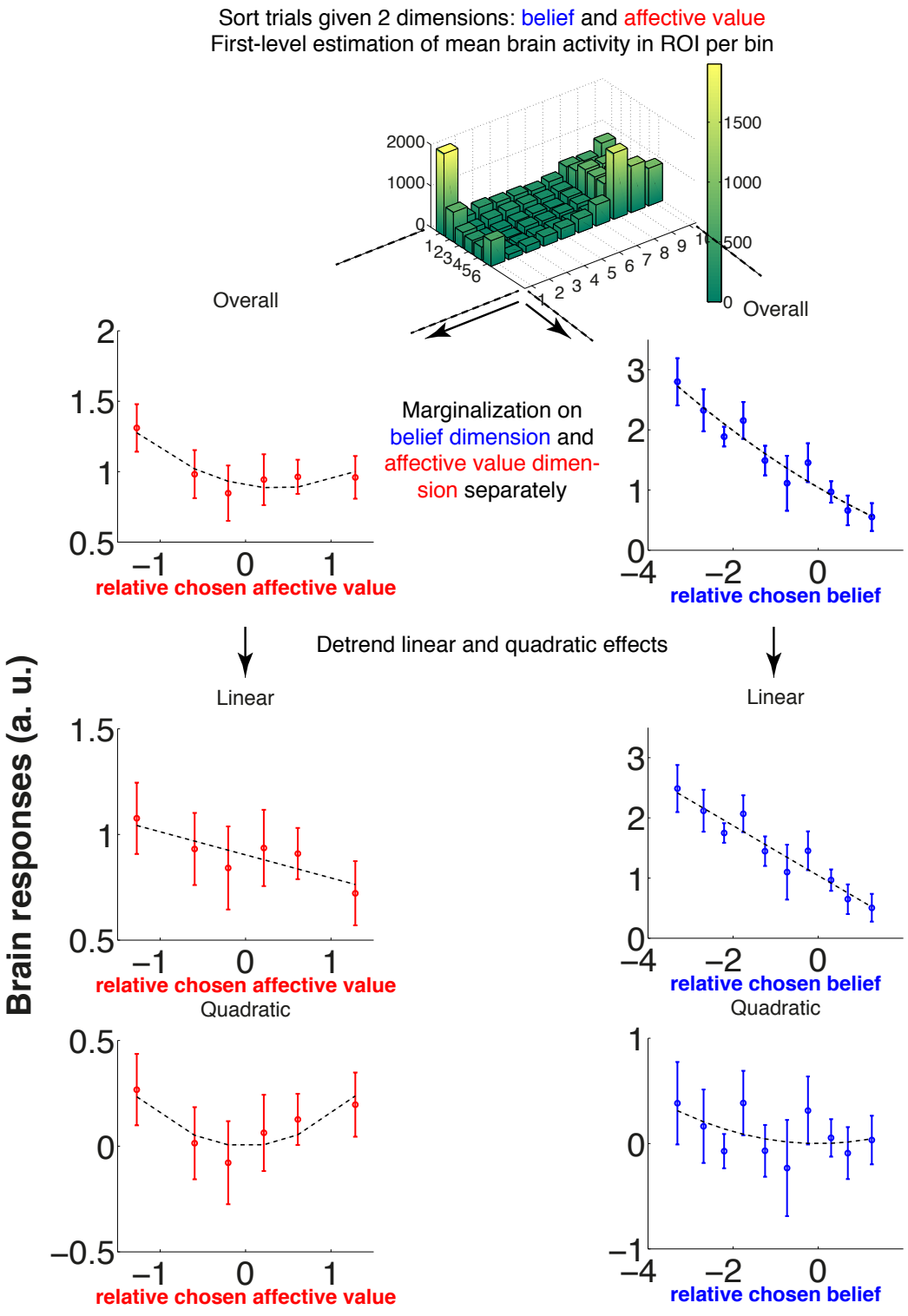


FIGURE 7.18: Schematic explanation of the 3D bins analysis. Each bar height represents the number of trials falling in that bin. Details are provided in the main text.

Chapter 8

Protocol B: Results

8.1 Behavior

In a probabilistic reversal-learning task, healthy human subjects had to decide between two shapes representing two underlying states, one of which was more frequently rewarded than the other one. The proposed rewards to gain for each shape were displayed before each choice. Crucially, we manipulated the reward distributions underlying each shape to dissociate beliefs from affective values, in three experimental conditions.

8.1.1 Learning curves

The upper panels learning curves in figure 8.1 show the choices proportion of the most frequently rewarded shape, plotted against trial number after a reversal. The choice frequency of the most frequently rewarded shape increased with after a contingencies reversal. First, we observed that after 5-10 trials after a reversal, subjects learnt which shape was the most frequently rewarded one. Participants reached an asymptotic behavioral performance (mean correct responses: 74.1%). After a reversal, subjects re-learnt from scratch to identify the most frequently rewarded shape, and eventually reached a probability-matching level (asymptote around 80%, figure 8.1).

The lower panel learning curves in figure 8.1 show the choices proportion of the shape with the highest expected value, plotted against trial number after a contingencies reversal. The choice frequency of the shape with highest expected value increased after a reversal. Differences at the asymptote were found between the three conditions. The asymptotic choice proportion of the highest expected value option at the plateau was higher on condition correlated than in condition random than in condition anti-correlated (paired t-test comparing plateau trials 9-16: condition anti-correlated vs. others: both

$p < 10^{-3}$, condition correlated vs. random at trend $p < 0.048$). If subjects were optimal, they would have chosen the shape with the highest expected value 100% of the time, in all conditions. This was not the case. As it is usually found in this kind of economic task, subjects were suboptimal (Trommershauser et al., 2008 [180]). Moreover, they departed from optimality differently in the three conditions. This provides evidence that subjects were sensitive to the informational value of proposed rewards that we differentially manipulated across conditions.

In the condition anti-correlated, the asymptotic level was higher for choice proportion of the most frequently rewarded shape than for the choice proportion of shape with highest expected value ($p < 10^{-4}$, paired t-test comparing plateau trials 9-16) (rightmost panels in figure 8.1). This result suggests that subjects favor accuracy, "being right", identifying the "correct" shape, rather than pure reward maximization. It seems that subjects acted as if they were trying to ignore rewarding values ; and focus only on probabilities (i.e. shapes). In the condition anti-correlated, they seemed to use the informational value carried by rewards and chose accordingly more often the shape associated with lower proposed rewards. Critically, the task was designed such that in all conditions, choosing the most frequently rewarded shape was always better on average.

8.1.2 Rewards

Figure 8.2 shows the proportion of time that each proposed reward was selected, regardless of the shape with which it was associated. As expected, in the condition correlated, the higher the proposed reward, the more often subjects selected it (Figure 8.2). Furthermore, in the condition random, rewards presented before choice were randomly drawn and carefully pseudo-randomized. So, if subjects were relying only on shapes, they should have chosen each proposed reward (2, 4, 6, 8, 10 Euros) with the same frequency on average. Essentially, if subjects were relying only on shapes, the orange line on figure 8.2 would have been flat. However, in condition random, we observed that subjects more often chose the highest of both proposed rewards (orange line in Figure 8.2), meaning that subjects were sensitive to the proposed reward magnitudes. The asymmetry between condition correlated and condition anti-correlated (yellow curve in Figure 8.2) also illustrates this sensitivity to reward magnitude i.e. affective value. The choice proportion of 6 Euros, the central value, did not vary across conditions. Therefore, subjects were sensitive to proposed rewards in their decisions. They did not base their decisions only on shapes. This behavioral measure will be used next to compare models.

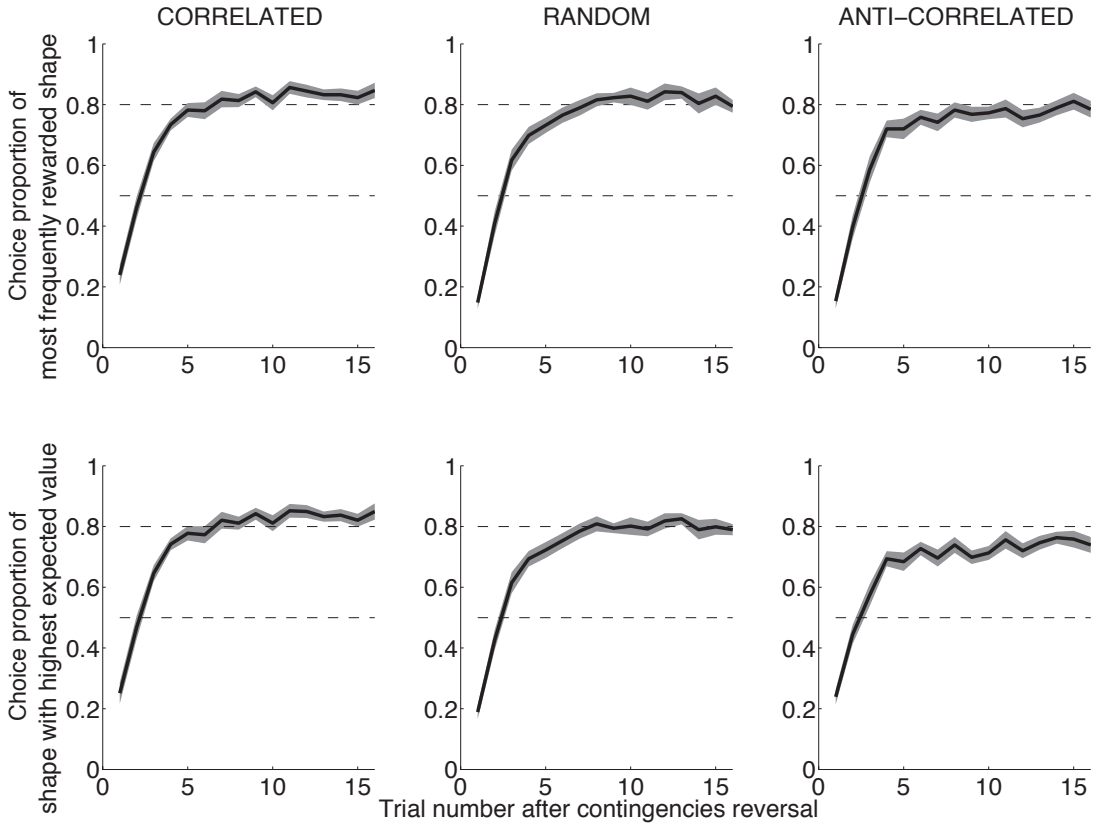


FIGURE 8.1: Learning curves. Upper panel: Choice proportion of most frequently rewarded shape, for the three experimental conditions. Left: condition correlated. Middle: condition random. Right: condition anti-correlated. Lower panel: Choice proportion of shape with highest expected value, for the three experimental conditions. Shaded area represents the standard error of the mean across subjects (average over reversals and over subjects (N = 22)).

8.1.3 Logistic regressions

We then investigated which possible variables could influence choice, using a logistic regression. This logistic regression was performed with a full variance analysis, meaning that each effect was truly specific to each regressor, while the common variance between regressors was removed. Therefore, it means that the order in which the regressors are presented does not matter. Moreover, all regressors were z-scored before entering the regression, meaning that the relative contribution of each can be compared. Figure 8.3 plots the subjects' propensity to choose the square shape according to various protocol variables. Critically, regressors were incrementally added; and likelihoods of each regression, taking into account the number of degrees of freedom, were compared.

- We show that the probability of the shape being rewarded influenced choice in all conditions (conditions anti-correlated and random: $p < 10^{-3}$, condition correlated

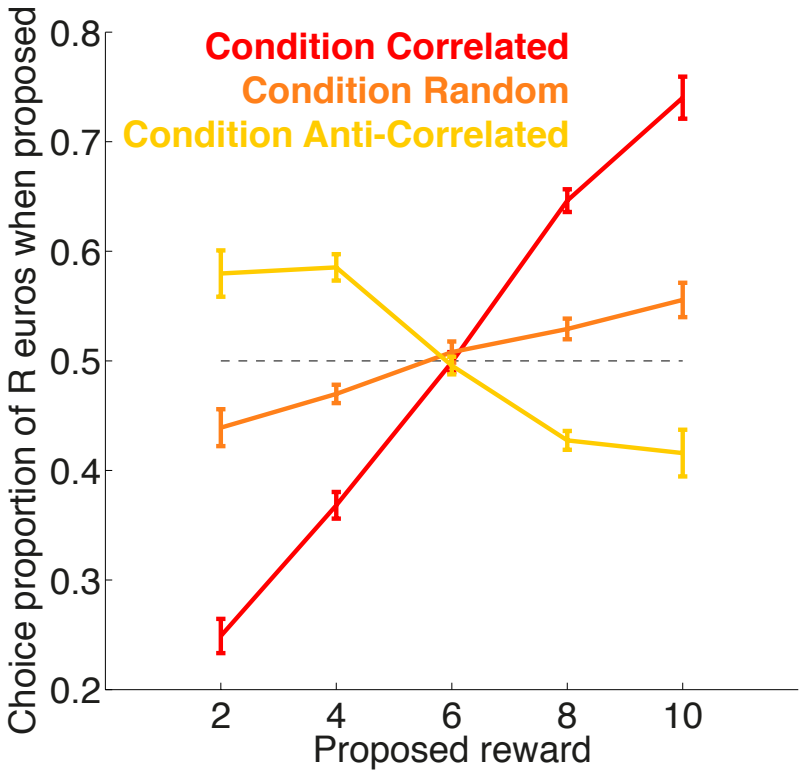


FIGURE 8.2: Choice proportion of R euros when proposed, regardless of shapes, for the three experimental conditions. Error bars represent the standard error of the mean (N = 22 subjects).

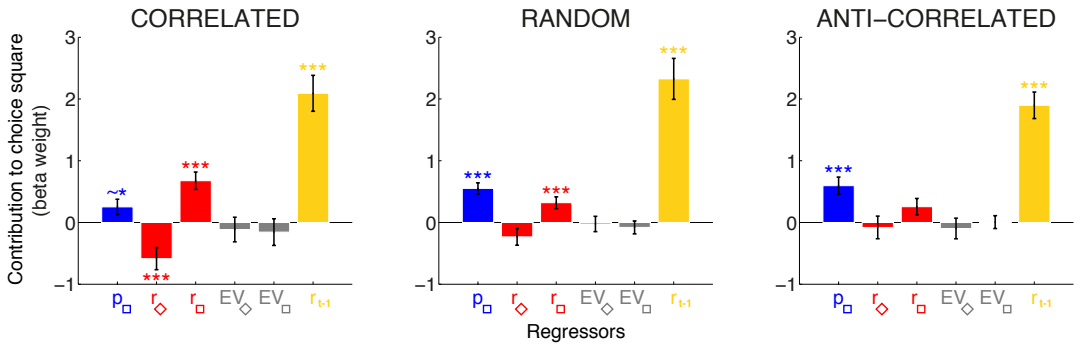


FIGURE 8.3: Logistic regression investigating the relative contribution of different protocol's variables to choice. Blue: probability associated with square shape; red: proposed rewards before decision; grey: expected value associated with each shape; yellow: reward received at previous trial. * $p < 0.05$, *** $p < 0.005$.

at trend: $p = 0.057$). This means that subjects were indeed looking for the most frequently rewarded shape.

- The proposed rewards before choice had a differential contribution across conditions. So, if the proposed rewards were processed only as affective values, i.e. 2 Euros is 2 Euros in any condition, we would not have observed a differential effect. This differential effect was related to subjects using the informational value conveyed by proposed rewards before choice.

  Importantly, no subject verbally reported detecting any differences between the three conditions, even when explicitly asked. However, they differentially used the proposed rewards:

  - Condition random: Rewards carried no information but subjects were slightly biased towards the highest proposed rewards. Red bars show the size of the baseline effect of pure affective value (Figure 8.3).

  - Condition correlated: Subjects were more driven towards high proposed rewards than in condition random. The difference between the two proposed rewards was more important in the condition correlated than in the condition random ($p = 0.029$).

  - Condition anti-correlated: Subjects were less driven towards high proposed rewards than in condition correlated ($p = 0.021$).

  Thus in conditions correlated and anti-correlated, subjects had the capacity to extract information contained in the proposed rewards to drive their choices. In other words, in the condition random, the proposed rewards had a small influence on choice, which represented a pure affective value baseline effect of proposed rewards before choice. A large effect was present in the condition correlated, which means that participants were more driven by proposed rewards when these proposed rewards were consistent with probabilities associated with shapes. Interestingly, in our pilot studies, in the condition anti-correlated, proposed rewards had an effect but in the opposite direction compared to the condition correlated, and of lower amplitude. Here, the same effect cumulated with the baseline pure affective value effect that was visible in the condition random resulted in no significant influence of proposed rewards for the condition anti-correlated.

  Overall, this result provides evidence that subjects extracted information from proposed the rewards before choice. The proposed rewards thus influenced subjects' decisions.

- Surprisingly, expected values (probabilities associated with shapes times proposed rewards) showed no significant contribution to choice (all $p > 0.4$). Participants

did not compute an expected value per se; they combined probability with proposed reward in a different manner, not explicitly calculating expected values. We checked statistically that adding the expected values did not significantly improve the logistic regression (paired t-tests on BIC: $p < 10^{-20}$).

- All regressions consistently showed a significant influence of the reward received at previous trial on the current choice (all $p < 10^{-5}$).

  Remark: The regression constant term was almost null except in the condition anti-correlated, in which more things might be going on (e.g. inhibitory processes).

When the reward received r_{t-1} at previous trial was replaced by coding a binary feedback (rewarded/not rewarded, x_{t-1}) at previous trial, we observed the same pattern of regressors contribution (Figure 8.4).

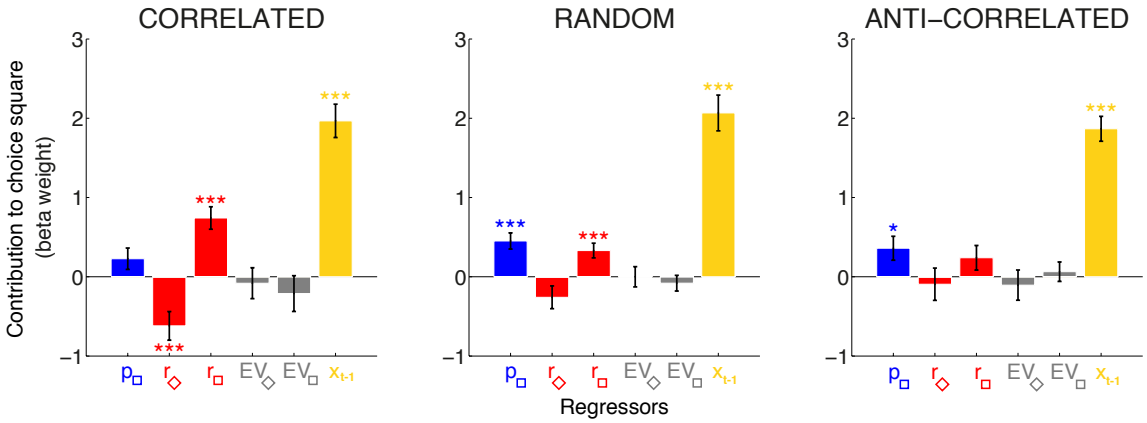


FIGURE 8.4: Logistic regression investigating the relative contribution of different protocol's variables to choice. Blue: probability associated with square shape; red: proposed rewards before decision; grey: expected value associated with each shape; yellow: binary feedback at previous trial i.e. rewarded/not rewarded.

A 7 by 3 ANOVA with factors REGRESSORS (probability, the two proposed rewards, the two expected values, the reward received at previous trial and the constant term) and CONDITION (correlated, random, anti-correlated) revealed a significant main effect of REGRESSORS ($F = 58.3$, $p < 0.001$), but no main effect of CONDITION ($F = 2.08$, $p = 0.15$) and no significant interaction between REGRESSORS and CONDITION ($F = 1.16$, $p = 0.34$). Although the interaction was not significant when including all regressors, post hoc tests revealed a significant difference across conditions regarding the difference between the two proposed rewards' effects (red bars in Figure 8.3, $F = 5.16$, $p = 0.02$). All other effects did not significantly vary across conditions (all $p > 0.1$).

Conclusion. Both probability associated with shape and proposed rewards influenced choice, but not in the form of a computation of an expected value. Importantly, it is the concomitance of the three experimental conditions, in which we modulated the reward distributions underlying each state, that permitted to dissociate the affective value from the information carried by proposed rewards.

8.1.4 Stay/Switch trials

The *stay* trials frequency, i.e. trials in which subjects chose the same shape as in the preceding trial is plotted as a function of the reward received at previous trial (Figure 8.5). Consistently with results from Protocol A, we observed a binary behavior, in which subjects tended to switch more after no reception of a reward, as compared to after reception of any other reward.

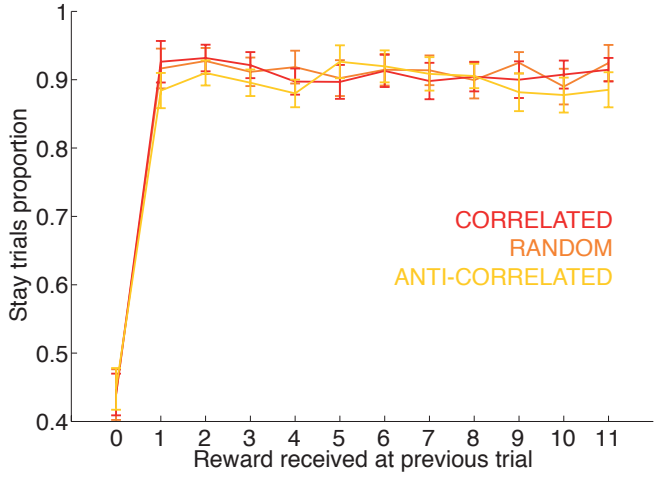


FIGURE 8.5: Stay trials frequency given reward received at previous trial.

8.1.5 Reaction times

Originally, we expected that reaction times will be higher in conditions correlated and anti-correlated compared to condition random, because in condition random, proposed rewards carried no informational value to process. In that sense, the conditions correlated and anti-correlated were richer. However, no significant difference was observed between the three conditions (paired t-tests, all $p > 0.5$). A slight but not significant increase in reaction times was observed following a reversal.

8.2 Modeling

8.2.1 First class of models

A first class of models that do not extract informational values from the proposed rewards before choice did not explain the subjects' behavior, as shown in model simulation plotted over subjects' average behavior (Figures 8.6 and 8.7). Mathematical description of Standard RL and Normalized RL models is provided in the Methods section and in Appendix D. These models were based on a continuous trial-by-trial update of option values and were blind to the task structure and reversals. Critically, the learning curve slope was smaller for the Standard RL model simulations as compared to subjects' behavior. It means that the Standard RL model adapted slower when a reversal occurred (left panel, Figure 8.6).

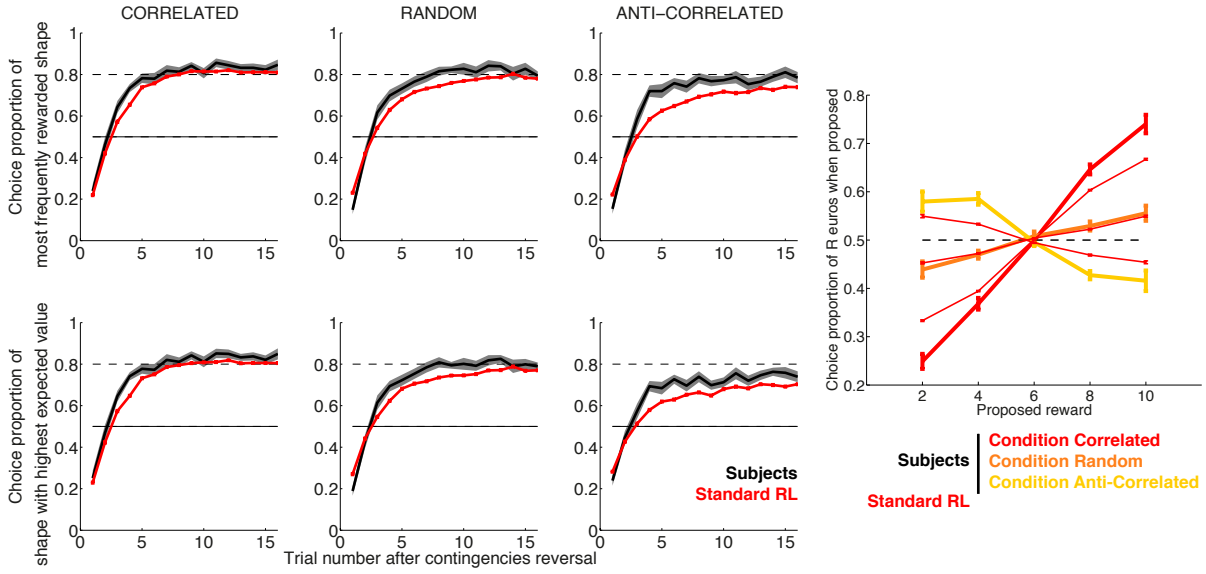


FIGURE 8.6: Simulations (N = 1000) of the Standard RL model (red) plotted over subjects' behavior (N = 22). Error bars represent the standard error of the mean.

The Normalized RL model corresponds to a RL model but with update of both chosen and unchosen option values. It hypothesizes that subjects make counterfactual inferences about the unchosen shape. Model details are provided in the Methods section. In fact, the normalized RL model was not able to capture subjects' behavior neither, especially in the condition anti-correlated, as shown in simulations in Figure 8.7. Even though the Normalized RL model performed above chance level, it remained less good than subjects in the condition anti-correlated, even at the asymptote (Figure 8.7). This result means that subjects did not make the inference that if they did not obtain a reward when choosing a shape, they would have obtained a reward should they have chosen the other

shape. This was not a trivial result because participants could have formed incorrect beliefs about the task in such binary choice settings.

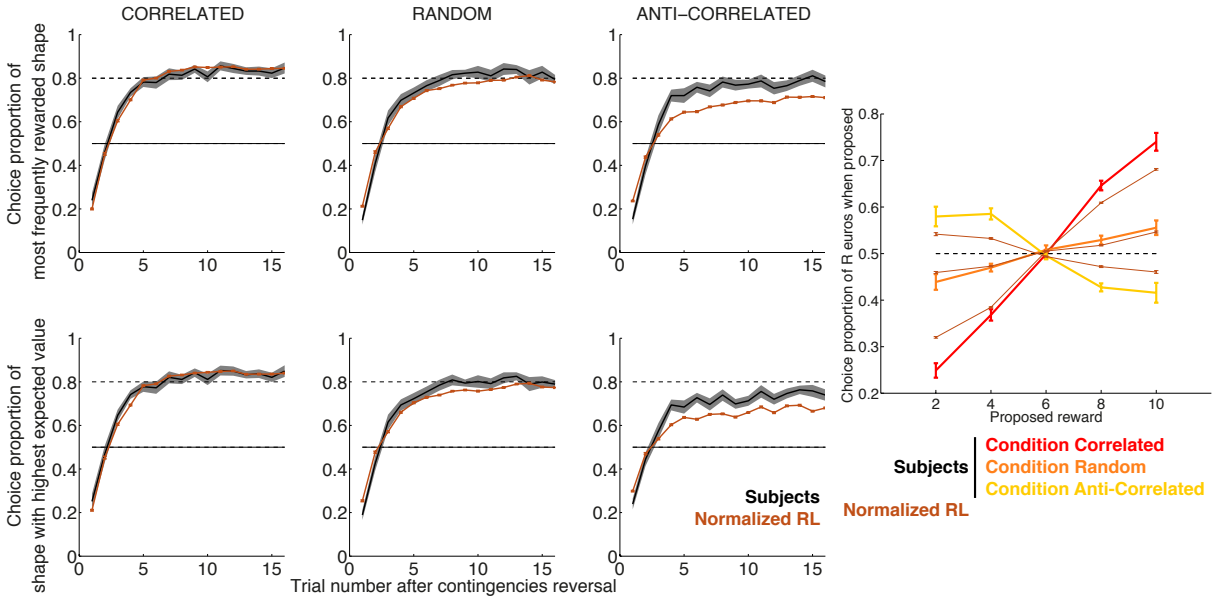


FIGURE 8.7: Simulations (N = 1000) of the Normalized RL model (brown) plotted over subjects' behavior (N = 22). Error bars represent the standard error of the mean.

Moreover, this first class of models (Standard RL, Normalized RL and others not shown) all explained the behavior significantly less well than many other alternative models, in terms of LLH, AIC and BIC as shown in Figure 8.13.

8.2.2 Second class of models

This second class of models stems from the Bayesian model of the task formally described in Appendix D which constitutes the statistically optimal behavior. More precisely, in this kind of economic decision-making task, optimal behavior consists in maximizing an expected value (probability of obtaining a reward multiplied by reward magnitude). This Bayesian model monitors a belief about how shapes map onto outcome contingencies (i.e. in this task, a belief about which shape is the most frequently rewarded one). A key feature of this second class of models is that the informational value from the proposed rewards presented before choice is extracted and used to update the belief before choice. Informational value conveyed by proposed rewards constitutes a likelihood in a Bayesian framework. An expected value is then computed for each shape: belief multiplied by proposed reward, and subjects (soft)maximize these expected values to choose. After choice, a binary feedback (win/lose) is extracted from outcome and used to update the

belief using Bayes rule. Simply put, if a positive outcome is received, the belief that the chosen shape is the most frequently rewarded one will increase.

The second class of models better fitted subjects' behavior than the first class of models (Figure 8.13 and simulations Figure 8.8), which was evidence that participants not only processed rewards for their pure affective value, but also extracted informational value from proposed rewards at decision time. But on the rightmost panel in Figure 8.8, we could see that the Bayesian model simulations did not reproduce the participants choices pattern in the condition anti-correlated (yellow curve in rightmost panel in Figure 8.8).

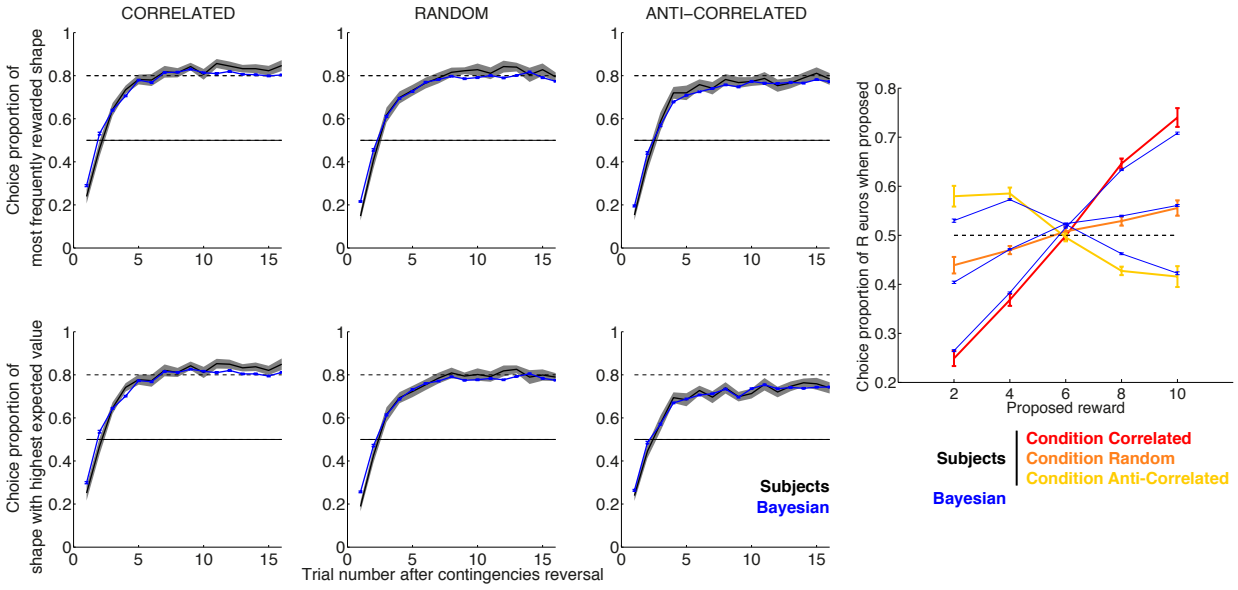


FIGURE 8.8: Simulations (N = 1000) of the Bayesian model (blue) plotted over subjects' behavior (N = 22). Error bars represent the standard error of the mean.

Nevertheless, subjects' apparent sub-optimality, as observed in learning curves (bottom panel in Figure 8.1), could be due to a misperception of actual probabilities and rewards, in the form of distortions in subjects' probability and reward representations. Therefore it could be that subjects did compute an expected value, but with distorted probability and distorted rewards; hence a seemingly suboptimal behavior. The distortions idea have been popularized in economy with the prospect theory (Kahneman and Tversky, 1979 [98]; Kahneman, 1984 [170]; Trommershauser et al., 2008 [180]). We fitted such a distortions model on our behavioral data, including all possible types of distortions for both probabilities and rewards (concave, convex, sigmoid, inverse sigmoid and absence of distortion), as formalized by Zhang and Maloney, 2012 [172]. This distortions model simulations displayed in Figure 8.9 well reproduced the pattern of subjects' behavior. In essence, the distortions model corresponds to the Bayesian model described by Behrens

and colleagues [162], although no volatility level is included since the volatility did not vary in our paradigm (see Chapter Methods).

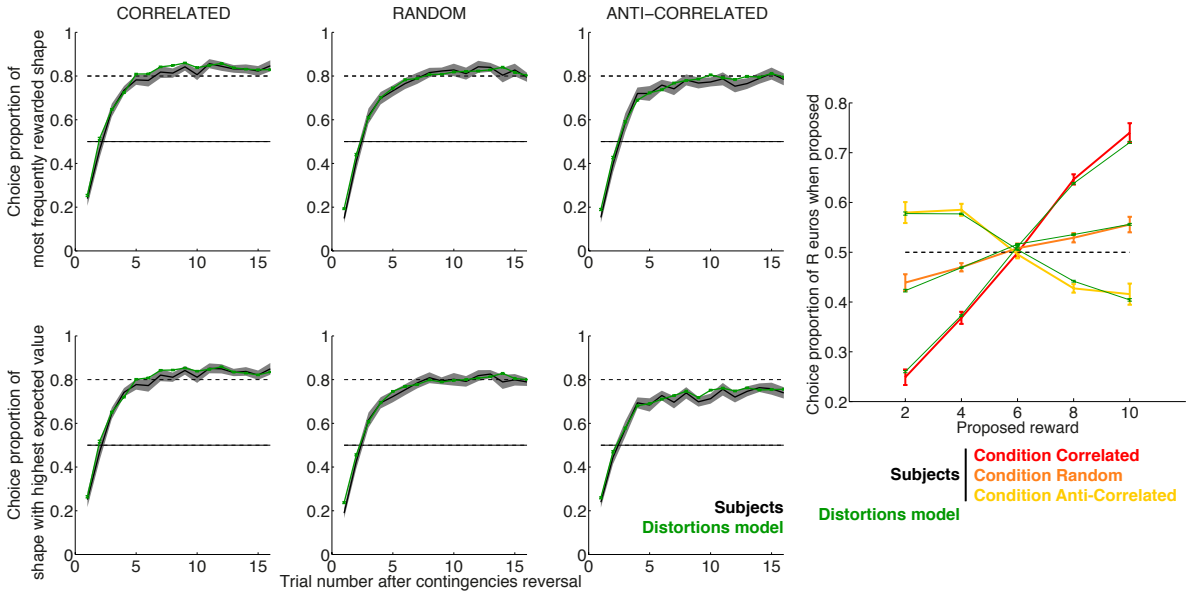


FIGURE 8.9: Simulations (N = 1000) of the distortions model (green) plotted over subjects' behavior (N = 22). Error bars represent the standard error of the mean.

The distortions model provided a good description of the data, and was able to capture the participants choices pattern. In particular, the distortions model captured the differential effect across conditions at the learning curves asymptote (Figure 8.9). However, this model raises two issues. First, it does not explain the psychological/cerebral source of these distortions. More precisely, it does not give a mechanistic explanation for why would subjects distort their probability and reward representations. Furthermore, the distortions that we obtained after fitting were contrary to what has been reported in the literature (Figure 8.10). On the one hand, we found no significant distortion on rewards, whereas rewards are usually flattened with increasing magnitude (Dehaene et al., 2009 [181]). On the other hand, we found a sigmoid distortion on probabilities, whereas probabilities deformation is usually an inverse sigmoid (Kahneman and Tversky, 1979 [98]: low probabilities generally tend to be overweighted; intuitive example being overestimating the probability of winning the lottery).

Instead, we propose that subjects do not compute expected values per se, as predicted by the Bayesian and distortions models, but combine beliefs about shape and proposed rewards in a different manner.

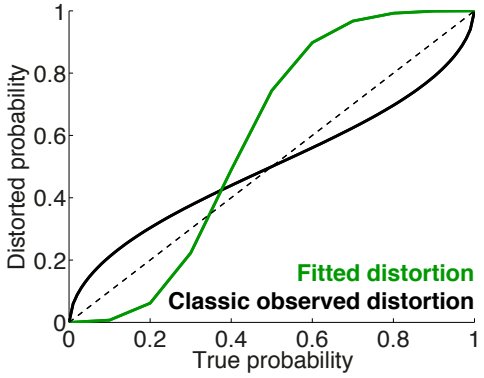


FIGURE 8.10: Fitted distortions in distortion model. Subjects tended to deform their probabilities estimates in a binary manner, opposed to what has been reported in the prospect theory.

8.2.3 Third class of models

The third class of models does not compute expected values but rather linearly combines beliefs and affective values to make a decision. In these models, choice was made over a mixture of beliefs (Bayesian inference system) and affective values (RL system), but not in the form of expected value computation. Further description of these models is provided in the methods section, let us focus on the main model in the third class: the mixed model.

According to the mixed model, a belief about how shapes map onto outcome distributions and an affective value are combined to make a decision (Figure 8.11). More precisely, the belief was a prior belief from the past. When subjects observed the proposed rewards before choice, the prior belief was revised by the informational value conveyed by proposed rewards (likelihood in a Bayesian framework). On the other hand, the affective value was the sum of a reinforcement value from previous trials and the proposed reward displayed before each choice. Subjects made a decision by combining these two systems. They subsequently observed an outcome. Given this outcome, the beliefs were updated by Bayes rule, while the affective values were updated by reinforcement learning (Figure 8.11).

The mixed model simulations revealed that it was a very good predictor of subjects' behavior, as exposed in Figure 8.12. The mixed model reproduced behavioral patterns of both learning curves and choice of proposed rewards.

In this mixed model, the fitted weight ω in the mixture of the two systems was in favor of the belief system (mean $\omega = 0.25$, Table 8.1). In other words, after averaging over subjects, choice was made on a mixture of 75% belief and 25% affective value. This

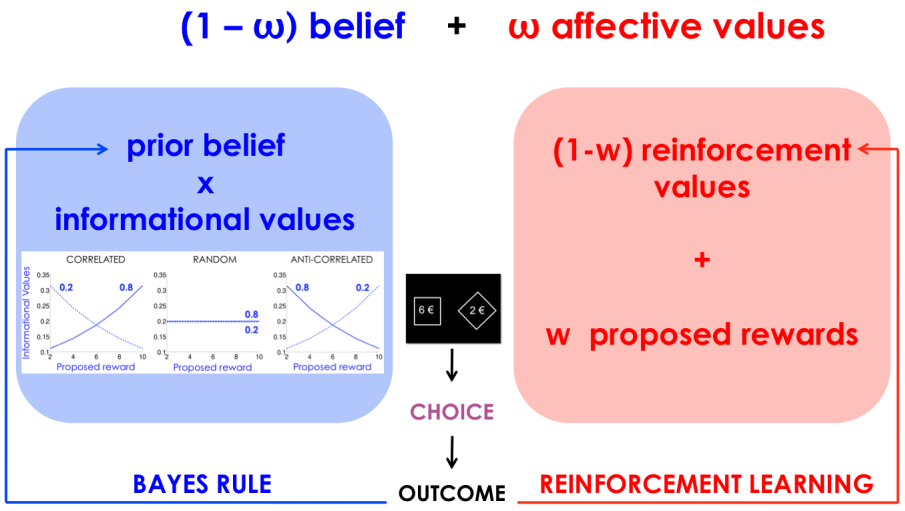


FIGURE 8.11: Schematic representation of the computations performed by the best-fitting mixed model. Details are provided in the main text.

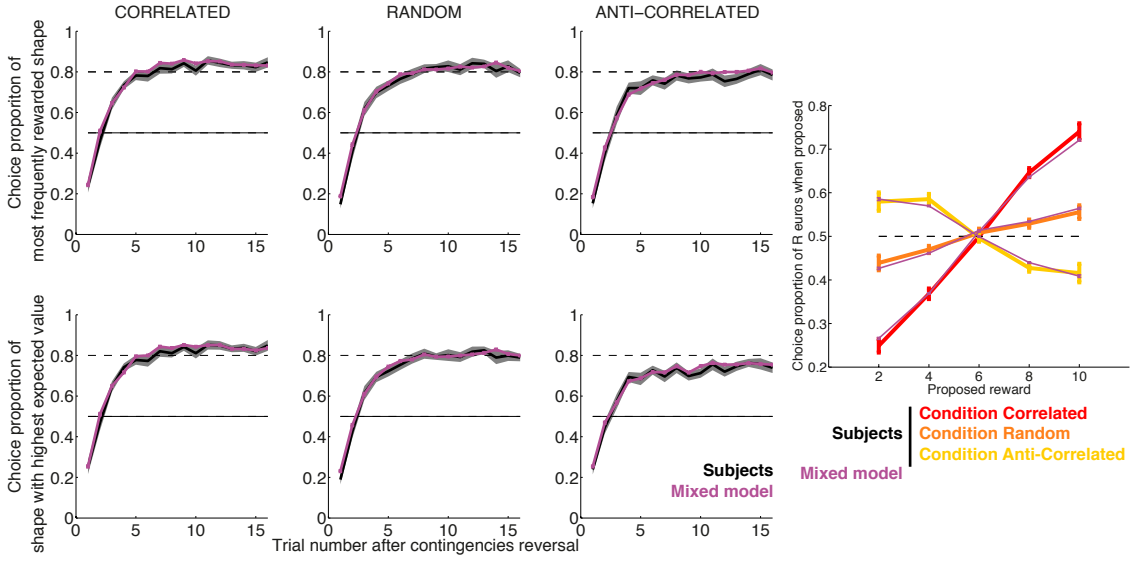


FIGURE 8.12: Simulations (N = 1000) of the best-fitting mixed model (purple) plotted over subjects' behavior (N = 22). Error bars represent the standard error of the mean.

means that the Bayesian inference system was predominant in decision, but including a slight role for the affective values from the RL system. Subjects tended to favor accuracy, "making the right choice", over pure reward maximization. Therefore, this third class of models can also be seen as Bayesian inference models monitoring beliefs but marginally biased by reinforcement affective values. However that we do not make any strong claim related to this parameter ω, because its value might be dependent on particular task settings.

Among this third class of models, we were able to rule out two other alternative models:

- Importantly, a particular sub-case of these mixed models with only a Bayesian system monitoring beliefs and without a reinforcement learning system was significantly a less good predictor of subjects' behavioral data ($p < 0.01$, paired t-test on BIC).

- Mixed models making a decision using a combination of (1) an expected value (belief times proposed rewards) from a Bayesian system and (2) reinforcement values from a RL system systematically fitted less well the behavior ($p < 0.00001$, paired t-test on BIC).

8.2.4 Model selection

In this section, we report quantitative criteria allowing for model comparison. However, we emphasize the importance of presenting qualitative models simulations, as in the above figures, to support a model relevance.

All classes of models were quantitatively compared in Figure 8.13, by summing LLH, BIC and AIC over all participants (fixed effects analysis). As described in the Methods section, the Akaike Information Criterion (AIC) is a measure of a relative statistical model's quality, trading-off goodness of fit (LLH) and parsimony (number of degrees of freedom, i.e. number of free parameters here). The Bayesian Information Criterion (BIC) also takes into account the number of free parameters, but additionally includes a factor penalizing for number of observations (i.e. number of trials here).

By looking only at behavioral simulations (Figure 8.9 and 8.12), we were not able arbitrate between our second class of models (distortions model, based on prospect theory) and our third class of models (mixed model, with no computation of expected value per se). Qualitatively, learning curves evolution was well reproduced in both classes of models. Quantitatively however, the mixed model interpretation was significantly favored, as shown in Figure 8.13. The mixed model better and more parsimoniously fitted subjects' behavioral data (random effects analysis, Figure 8.14, paired t-tests, LLH: $p = 0.058$,

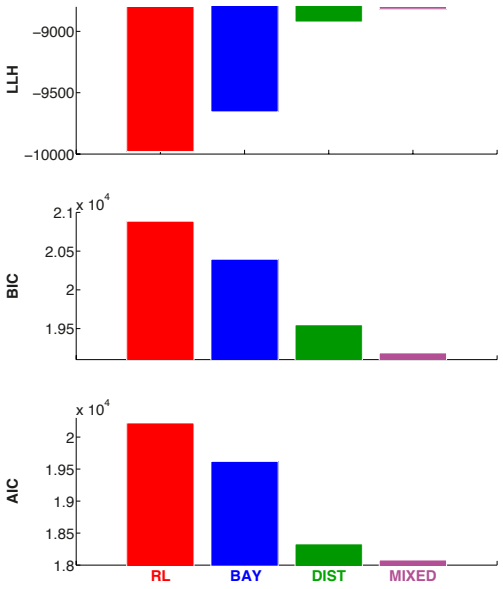


FIGURE 8.13: Full model comparison shows that the mixed model best fitted the subjects' behavioral data (fixed effects analysis).

BIC: $p < 0.005$, AIC: $p < 0.05$). Moreover, our mixed model provides a mechanistic explanation of computations underlying choice, not only a psychological description.

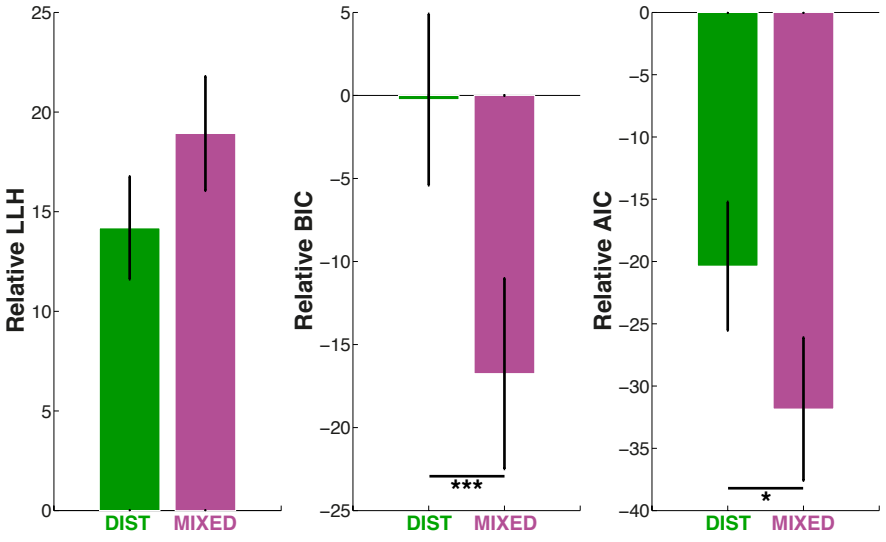


FIGURE 8.14: Comparison of the two best-fitting models in terms of relative Log-likelihood, Bayesian Information Criterion and Akaike Information Criterion, as compared to a baseline model consisting of only a belief system (random effects analysis). *** $p < 0.005$, * $p < 0.05$.

However, the distortions model and the mixed model are not necessarily contradictory accounts. Indeed, the sigmoid distortion observed on probabilities in the distortions

model means that subjects had a binary perception of probabilities (a "good" shape and a "bad" shape). This was fully consistent with the predominance of the belief system in the mixture in the mixed model. This indicates that the belief about how shapes mapped onto reward distributions mattered more than rewards (affective value) in decision.

8.2.5 Best-fitting mixed model parameters

Table 8.1 shows the mean and standard error of each free parameter adjusted to subjects' behavioral data. Further description and role of each parameter was provided in the Methods section. Overall, average fitted parameters were coherent with the task design parameters. Volatility was slightly overestimated (average: 0.16) compared to its real value (reversals frequency: 0.05). This result has been consistently observed across different paradigms in our team. A possible explanation could be that this parameter could be a second order estimate. Consequently, subjects had more difficulty estimating it, or perceived the environment as more volatile or more uncertain than it was in reality.

Parameters	**Description**	**Mean**	**S.E.M.**
β	Inverse temperature in softmax	44.2	9.4
volatility	Volatility in Bayesian system	0.16	0.03
ϵ	Lapses rate in softmax	0.02	0.01
q	Probability of obtaining a reward (0.8 in design)	0.74	0.03
γ *correlated*	Slope of the reward distribution in condition correlated	0.041	0.048
γ *random*	Slope of the reward distribution in condition random	-0.005	0.003
γ *anti-correl.*	Slope of the reward distribution in condition anti-corr.	-0.095	0.060
learning rate	Learning rate in RL system	0.72	0.08
w	Bias towards the current proposed reward in RL	0.28	0.07
ω	Weight between the two systems in decision	0.25	0.05

TABLE 8.1: Best-fitting mixed model parameters. Mean and standard error of the mean (S.E.M.) across subjects (N = 22) are provided. Weight ω corresponds to the weight of the affective values in decision.

In addition, we observed that the average fitted weight favored the belief over the affective values in the decision. Indeed, we found after fitting $\omega = 0.25$ on average, with $\omega < 0.1$ for more than a third of subjects. The actual distribution of the weight parameter within the group (N = 22 subjects) is provided in Figure 8.15.

Although subjects probably used various strategies to solve the task, we did not observe a multimodal distribution with clear distinct groups of subjects emerging. The mean of the weight parameter thus seems representative of the group. The predominance of the belief system appears robust across subjects. Nonetheless, the relative contribution of each system in the mixture must not be over-interpreted. It probably depends on the

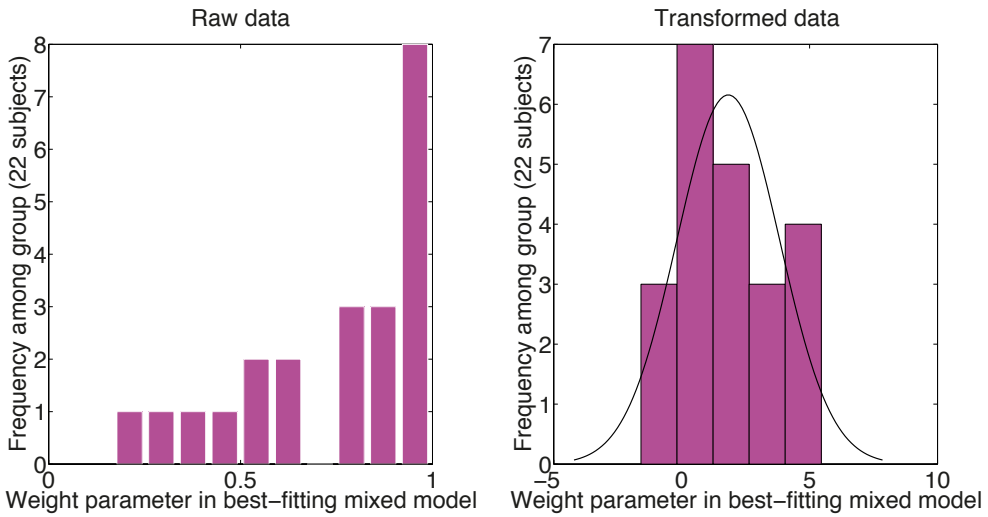


FIGURE 8.15: Distribution of the fitted weight parameter within the group. Left panel represents raw data in the form of 1-ω i.e. the contribution of the belief system. Right panel: same data after log transformation, with a gaussian fit.

particular task design and, more generally, on the current ecological situation at hand (e.g. volatility, gains at stake).

Importantly, when we fitted three weight parameters for the three conditions instead of one, the model's BIC was not improved ($p = 0.26$). Moreover, the three fitted weights were not different from each other (all $p > 0.08$).

Finally, fits of the inverse temperature β and lapses rate ϵ of the model's ϵ-softmax resulted in high values for β and low values for ϵ (Table 8.1). This provides further evidence that our best-fitting mixed model had a good explanatory power.

8.2.6 Informational Values

The only free parameter allowed to vary across the three experimental condition was γ, the slope of the reward distributions. Importantly, these free parameters (γ *correlated*, γ *random*, γ *anti-correl.*) fitted on the behavioral data were able to capture the actual reward distributions tendency imposed by the experimental design; with γ *correlated* being positive, γ *random* being close to zero and γ *anti-correl.* being negative (Figure 8.16 and Table 8.1).

γ *anti-correlated* was significantly different from γ *correl.* and from γ *random* (both $p < 0.05$), when parameters were fitted on the second half of trials of each session. When the fit was done including all trials, they remained marginally significant (γ *correlated* vs. γ *anti-correl.*, $p = 0.06$). This just means that it took more time than the training before

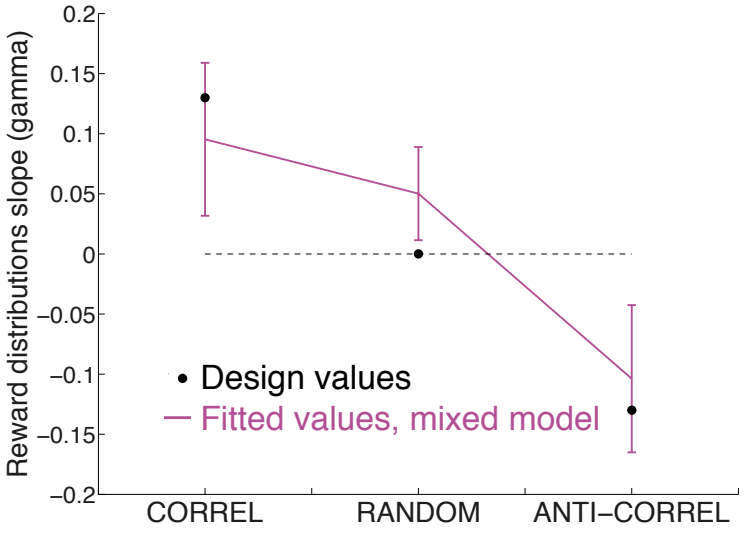


FIGURE 8.16: Fitted values of the free parameter γ for each experimental condition: γ *correlated*, γ *random* and γ *anti-correl.*. Fitted values were able to capture actual design values.

each session to subjects to infer the reward distributions shape in each experimental condition. In addition, if the informational value update before decision in the Bayesian system was removed from the model, we observed a significant qualitative difference between the subjects' learning curves and the model's simulations (Figure 8.17). This was evidence that the informational value (likelihood) update was a crucial step in the model, even if the effect sizes were small in the fitted parameters (Figure 8.16).

Moreover, if we imposed the same parameter for the three conditions instead of three different free parameters, or if we imposed γ *correlated* = γ *anti-correl.* (two parameters instead of three), the model fitted significantly less well.

8.2.7 Conclusion

According to the mixed model, decision-making appears to result from a linear mixture of two independent systems, rather than an explicit computation of expected values (multiplicative). Beliefs (Bayesian system) and affective values (RL system) were combined to make a decision. Converging evidence showed that both the belief and the affective values systems contribute to choice. Nevertheless, both the distortions model and the mixed model have a similar explanatory power, with a similar LLH. The decision values in both models are close, but their internal variables differ. The distortions model remains a good descriptor of the data but had more free parameters. The mixed model is simpler. Therefore, it will be used to examine brain activations.

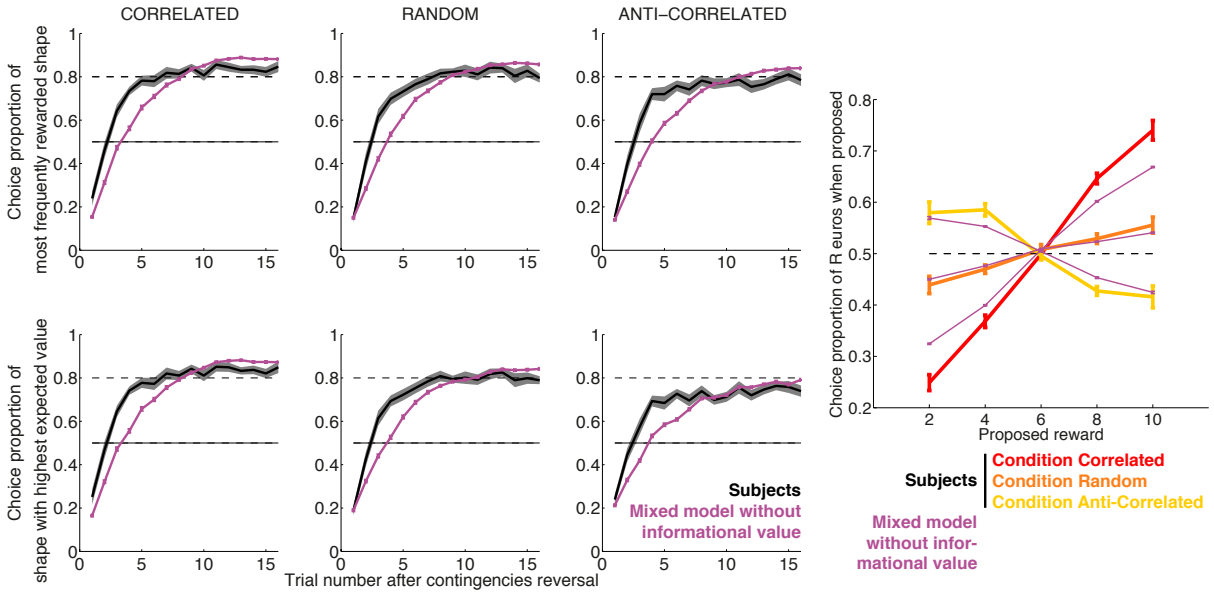


FIGURE 8.17: Simulations (N = 1000) of the mixed model with removal of the informational value update in the Bayesian system (purple) plotted over subjects' behavior (N = 22). The model does not capture subjects' behavior, showing that the update of the belief by the informational value of proposed rewards is a critical part of the model. Error bars represent the standard error of the mean.

8.3 Neuroimaging

We then used functional MRI to investigate how the belief system and the affective values system interact in the brain. Specifically, we examined whether the belief system and the affective value system had distinct neural bases. Principally, we will focus on describing activations in the frontal lobes. BOLD signal was regressed against choice-dependent (linear) and choice-independent (quadratic) representations of both beliefs and affective values. Critically, the quadratic effect was orthogonalized on the linear effect (Further details are provided in the Methods section). All the following second-level parametric maps were thresholded at $p < 0.005$ uncorrected, and for a cluster size of minimum 10 voxels. This threshold corresponds to a correction for the whole frontal cortex, region with a strong a priori, which consists of about a third of the brain. We describe below activations at stimulus onset, covering decision time window.

8.3.1 GLM1: Decision Values

We first examined the decision values neural correlates (Figure 8.18).

FIGURE 8.18: Best-fitting mixed model includes both a belief system and an affective value system that are combined to make a decision.

8.3.1.1 Choice-dependent effects

We observed choice-dependent effects in two main regions within medial prefrontal cortex.

Positive linear effects in vmPFC. We found a positive linear effect in ventromedial prefrontal cortex (vmPFC), extending into anterior PFC BA 10 (peak voxel at MNI coordinates [-9, 53, -5], T = 9.21). In others words, vmPFC activity correlated positively with the relative chosen decision value (Figure 8.19). Such a positive linear effect means that vmPFC activity increased when relative chosen decision value increased, which could reflect expectations associated with the chosen shape (action outcome expectation).

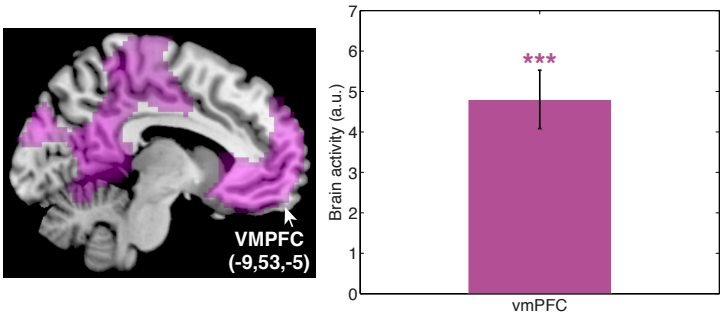


FIGURE 8.19: Positive linear effect of decision values in vmPFC. Left panel: parametric map thresholded at $p < 0.005$, $c > 10$ voxels, MNI peak voxel coordinates are indicated in brackets. Right panel: Effect size. Error bars correspond to the standard error of the mean, 21 subjects. a.u.: arbitrary units.

Negative linear effects in MCC. We found a negative linear effect in midcingulate cortex (MCC/dACC) (peak voxel at MNI coordinates [9, 20, 46], T = 11.60), including voxels falling into SMA. MCC activity correlated negatively with the relative chosen decision value (Figure 8.20). In other words, MCC activity decreased when relative chosen decision value increased. This pattern could reflect encoding of the unchosen decision value.

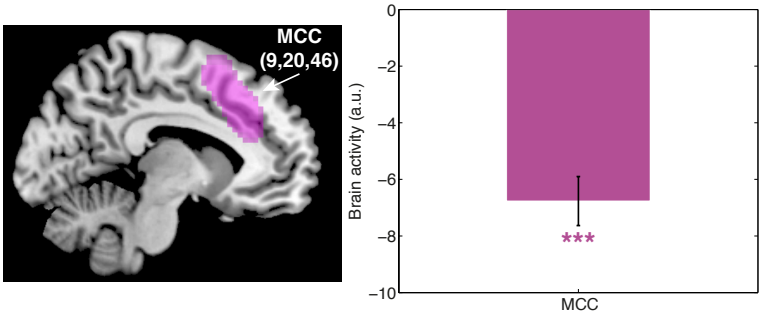


FIGURE 8.20: Negative linear effect of decision values in MCC. Left panel: parametric map thresholded at $p < 0.005$, $c > 10$ voxels, MNI peak voxel coordinates are indicated in brackets. Right panel: Effect size. Error bars correspond to the standard error of the mean, 21 subjects. a.u.: arbitrary units.

Critically, both ventral (vmPFC) and dorsal (MCC) parts of medial prefrontal cortex thus exhibited choice-dependent representations of decision values. These linear representations being signed by choice, they consist of post-choice representations (or concomitant to choice). Importantly, we reproduced here the classic effect of value encoding/chosen value expectation usually found in vmPFC in a number of studies (Lebreton et al., 2009 [49]; Plassmann et al. 2007 [53]; Chib et al., 2009 [52]).

8.3.1.2 Choice-independent effects

Positive quadratic effects in vmPFC and MCC. We showed choice-independent representations in both MCC (posterior, MNI peak coordinates: [0, 14, 37], T = 3.72) and vmPFC (large cluster in with voxels in posterior and medial orbital gyri and in anterior cingulate cortex; two main peaks: [0, 11, -14], T = 4.65 and [0, 41, 1], T = 4.22) as presented in Figure 8.21. Such a positive quadratic effect means that MCC and vmPFC activity increased when both chosen and unchosen decision values were far from each other, and activity was less intense when both chosen and unchosen decision values were close to each other.

Such a U-shaped pattern could reflect pre-choice preferences encoding, unsigned by choice. Alternatively, it could reflect an encoding of confidence about choice, i.e. more activity when the two decision values were far from each other so when the choice was easier. More precisely, the reasoning behind a confidence signal interpretation would be the following. When the two values, chosen and unchosen are far from each other, subjects would be more certain about their choice (be it certain they made the right choice or certain they made the wrong choice). On the contrary, when the two values are close to each other, subjects would be quite uncertain about their choice, and consequently

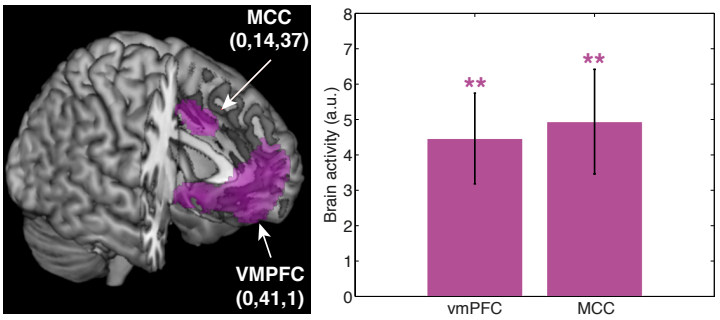


FIGURE 8.21: Positive quadratic effect of decision values in vmPFC and MCC. Left panel: parametric map thresholded at $p < 0.005$, $c > 10$ voxels, MNI peak voxel coordinates are indicated in brackets. Right panel: Effect sizes. Error bars correspond to the standard error of the mean, 21 subjects. a.u.: arbitrary units.

confidence should be low. In other terms, in that case, confidence would correspond to the absolute (i.e. quadratic) difference between chosen and unchosen values. However, further data shown in the next section rather support an interpretation in terms of choice-independent preferences than in terms of confidence.

Negative quadratic effect in lateral PFC. A brain region showing a negative quadratic effect corresponded to a region that was more activated when both decision values chosen and unchosen were close to each other, so when choice was more difficult. When we regressed decision values, we found a negative quadratic effect in lateral prefrontal cortex, bilaterally (Figure 8.22). The right activation (MNI peak coordinates: [39,53,13], T = 5.62) covered a large region from the dorsal part up to frontopolar cortex, while the left activation (MNI peak coordinates: [-39,59,4], T = 3.72) was smaller. Thus, lateral PFC activity increased when both chosen and unchosen decision values were close to each other, i.e. when action selection was more difficult. Such a pattern characterizes a region performing action selection.

8.3.2 GLM2: Dissociation belief system/affective values system

We then tested whether there were dissociated neural correlates for the belief system and the affective values system (Figure 8.18). Crucially, the analysis was performed in unique variance, meaning that the shared variance was removed and the remaining observed variance in the following second-level maps was truly selective of each system.

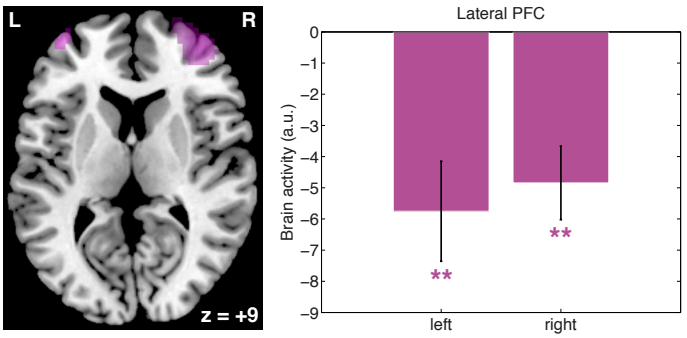


FIGURE 8.22: Negative quadratic effect of decision values in lateral PFC left (-39,59,4) and right (39,53,13), MNI coordinates. Left panel: parametric map thresholded at $p < 0.005$, $c > 10$ voxels. Right panel: Effect sizes. Error bars correspond to the standard error of the mean, 21 subjects. a.u.: arbitrary units.

8.3.2.1 Choice-dependent effects

Positive linear effects. As expected from previous studies (Plassmann et al. 2007 [53], Hampton et al., 2006 [56]), vmPFC activity correlated positively with both the relative chosen belief (MNI peak coordinates: [-9, 50, -5], T = 7.09) and the relative chosen affective value (MNI peak coordinates: [6, 26, -8], T = 5.52). Indeed, vmPFC activity increased when chosen belief and chose affective value increased, which reflected expectations associated with the chosen shape (action outcome expectation in terms of beliefs and in terms of affective values) (Figure 8.23).

In addition, for affective values, a positive linear effect was found in bilateral hippocampus (left: MNI peak coordinates: [-33, -16, -14], T = 8.20, right: MNI peak coordinates: [30, -19, -11], T = 4.68), suggesting that hippocampus is also involved in representing the expectations associated with chosen shape. This result is in line with a study by Lebreton and colleagues (Lebreton et al., 2013 [182]) showing that the hippocampus is involved in the valuation of imagined expected rewards. In our case, the likely interpretation is that once an action was chosen, subjects were anticipating the associated outcome.

Negative linear effects. As shown in Figure 8.24, MCC activity linearly varied negatively with both relative chosen belief (MNI peak coordinates: [3, 17, 52], T = 6.80) and relative chosen affective value (MNI peak coordinates: [3, 23, 46], T = 5.15). In other terms, MCC activity decreased when chosen belief and chosen affective value increased. The identified clusters extend dorsally, also including voxels of SMA and of dmPFC (BA9). The insula also correlated negatively with both relative chosen belief (left: MNI peak coordinates in inferior frontal gyrus: [-30, 23, -2], T = 6.17 and right: MNI peak coordinates in BA47: [33, 26, 12], T = 7.31) and relative chosen affective value (left:

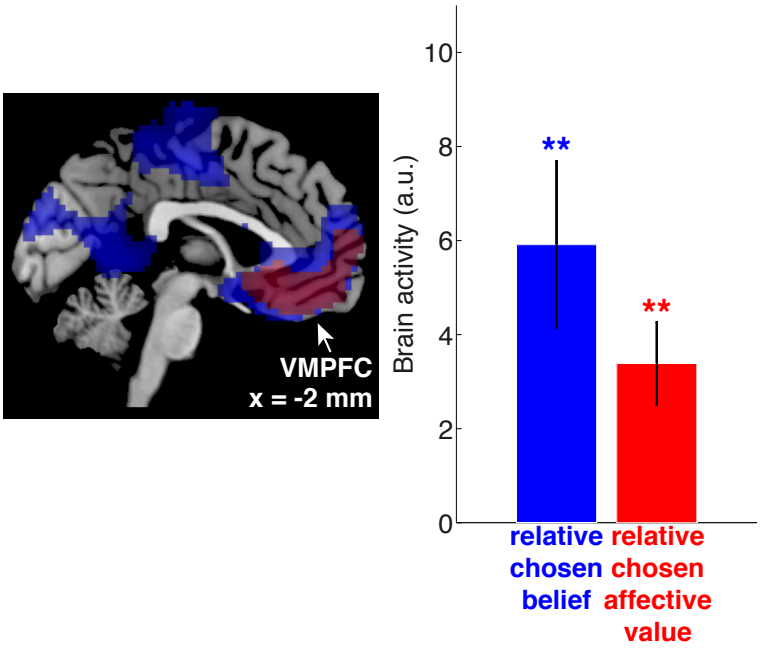


FIGURE 8.23: Positive linear effects in vmPFC for both relative chosen belief and relative chosen affective value.
Left panel: Axial brain slices with activations (thresholded at $p < 0.005$, voxel-wise, uncorrected) corresponding to relative chosen belief (blue) and relative chosen affective value (red) superimposed on anatomical template. x is brain slice MNI coordinate. Right panel: Effect sizes for relative chosen belief (blue) and relative chosen affective value (red) averaged over voxels from a sphere of radius 13 mm centered on the activation peak. a.u. arbitrary units. Error bars correspond to s.e.m across subjects. **$p < 0.01$.

MNI peak coordinates: [-30, 20,1], T = 4.07 and right: MNI peak coordinates: [33, 20, -2], T = 5.17). In addition, frontopolar cortex correlated negatively with both relative chosen belief (MNI peak coordinates: [-39, 59, -5], T = 3.01) and relative chosen affective value (left: MNI peak coordinates: [-36, 53, 10], T = 3.60 and right: MNI peak coordinates: [30, 56, 19], T = 4.03).

Therefore, no dissociation was found in vmPFC and MCC between beliefs system and affective values system regarding choice-dependent linear effects. In addition, we systematically observed stronger activations with the beliefs, which is consistent with our mixed model fitting showing that beliefs contributed more in the mixture.

However, we found a dissociation regarding choice-independent effects.

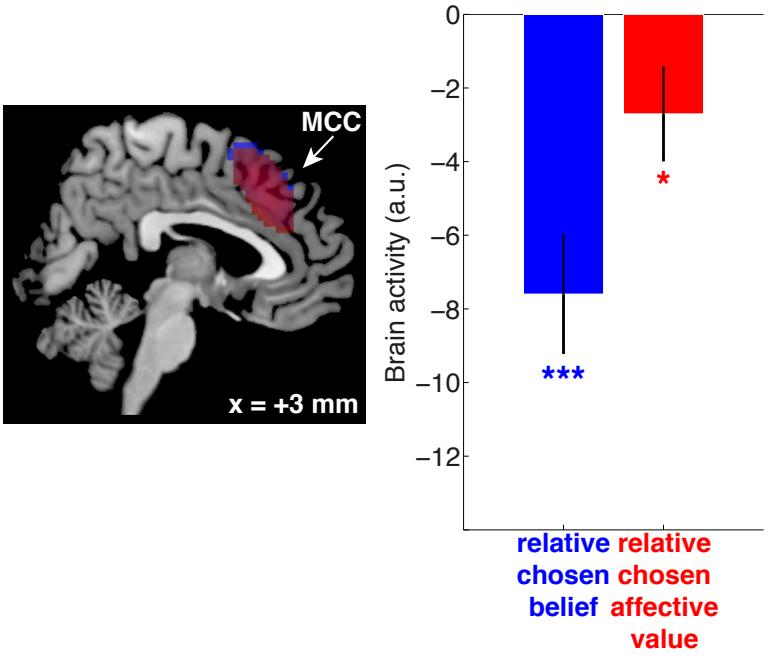


FIGURE 8.24: Negative linear effects in MCC for both relative chosen belief and relative chosen affective value.
Left panel: Axial brain slices with activations (thresholded at $p < 0.005$, voxel-wise, uncorrected) corresponding to relative chosen belief (blue) and relative chosen affective value (red) superimposed on anatomical template. x is brain slice MNI coordinate. Right panel: Effect sizes for relative chosen belief (blue) and relative chosen affective value (red) averaged over voxels from a sphere of radius 13 mm centered on the activation peak. a.u. arbitrary units. Error bars correspond to s.e.m across subjects. **$p < 0.01$.

8.3.2.2 Choice-independent effects

Positive quadratic effects. We observed a double dissociation between MCC and vmPFC. Surprisingly, vmPFC was specific to beliefs while MCC was specific to affective values.

- vmPFC (MNI peak coordinates: [-3, 44, -17], T = 4.34) activity increased when both beliefs were far from each other, and decreased when they were close to each other. By contrast, affective values did not modulate vmPFC activity regarding positive quadratic effects. Although negative results should be interpreted with much caution, we found no other frontal region for the positive quadratic effect of relative chosen belief.

- MCC (MNI peak coordinates: [0, 26, 40], T = 4.14) activity increased when both affective values were far from each other, and decreased when they were close to each other. By contrast, beliefs did not modulate MCC activity regarding positive

quadratic effects (Figure 8.25). Although negative results should be interpreted with much caution, we found no other frontal region for the positive quadratic effect of relative chosen affective value. The MCC cluster identified here was rather dorsal, i.e. in the midcingulate gyrus (dACC), with part of the voxels falling in BA32. Furthermore, certain human subjects have an additional cingulate sulcus which is dorsal to the first one and named paracingulate sulcus (Petrides et al., 2012 [15]; Amiez et al., 2013 [16]). I must acknowledge that I have not taken into account the variable presence of this paracingulate sulcus across subjects when performing fMRI analysis. fMRI activations were averaged over subjects and mapped onto a template brain.

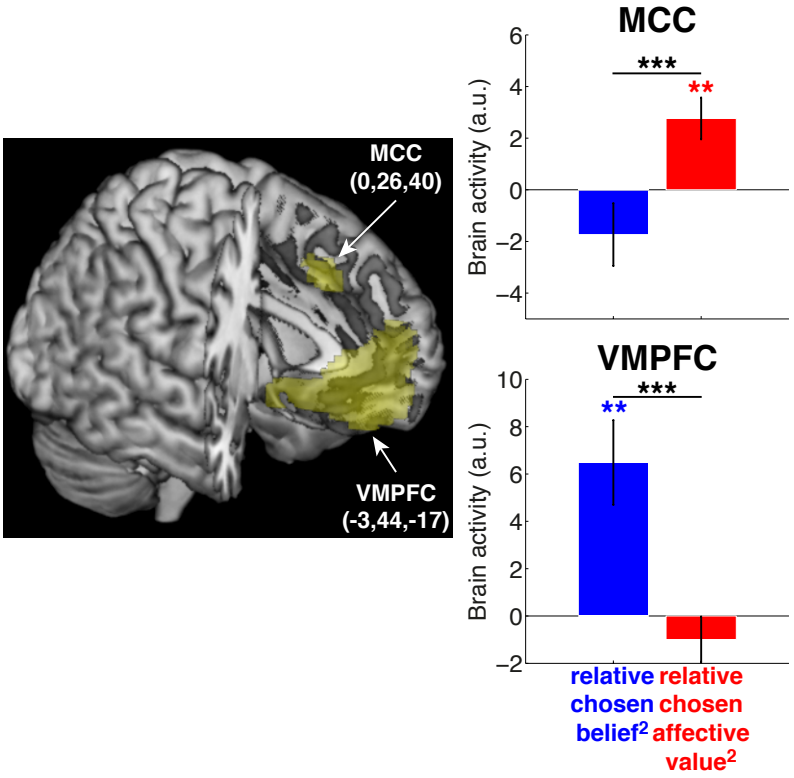


FIGURE 8.25: Double-dissociation MCC/VMPFC regarding choice-independent (quadratic) brain activations.
Left panel: 3D rendering of parametric brain activations correlating with relative chosen belief^2 (blue) and relative chosen affective value^2 (red) thresholded at $p < 0.005$ (voxel-wise, uncorrected). Coordinates (x,y,z) of activation peaks are from MNI space. Right panel: Effect sizes for relative chosen belief^2 (blue) and relative chosen affective value^2 (red) averaged over voxels from a sphere of radius 13 mm centered on the activation peak. a.u. arbitrary units. Error bars correspond to s.e.m. across subjects (N = 21). $^{**}p < 0.01$.

Besides, we observed a slight intra-individual correlation between the effect size (regression coefficient) of the positive quadratic effect of relative chosen belief in the vmPFC

and the weight attributed to the belief in the decision, as measured by the parameter ω from our best-fitting mixed model (Figure 8.26).

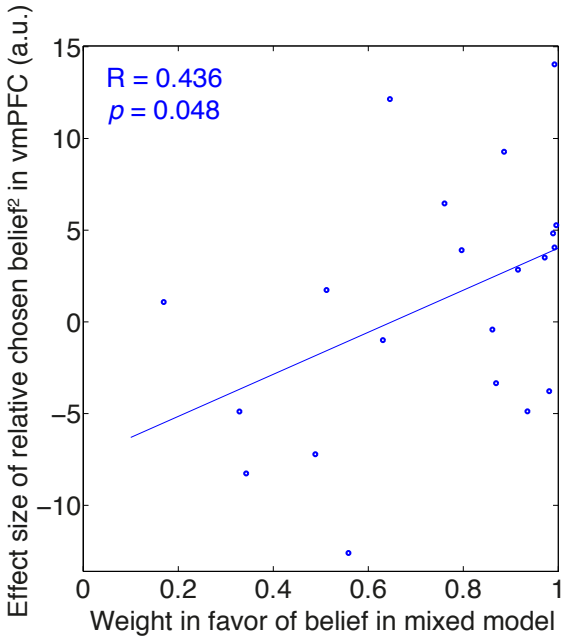


FIGURE 8.26: Scatterplot of the correlation between the effect size for the relative chosen belief^2 in vmPFC and the fitted weight parameter attributed to the belief in the decision from the best-fitting mixed model. Each dot corresponds to one subject. vmPFC ROI is defined from the second-level parametric map of the relative chosen belief positive linear effect.

Therefore, MCC encoded unsigned preferences in terms of affective values, whereas vmPFC encoded unsigned preferences in terms of beliefs, irrespective of choice. The implication of MCC in representing choice-independent affective values is in line with a lesion study in monkeys showing that the cingulate was necessary to integrate reinforcement values of food rewards over time, maintaining action/outcome history (not just detect errors in a single trial) (Kennerley et al., 2006 [183]).

In our protocol, since there was a double dissociation, it is more likely that the observed quadratic effects reflect unsigned choice-independent preferences rather than a confidence signal about choice. Indeed, confidence should be a post-choice global signal, not dissociated. In other words, we would expect a confidence signal to be signed by choice, occurring concomitantly or after choice. However, the positive quadratic effect observed in vmPFC for the belief could contribute upstream to the construction of a confidence signal. This point will be further discussed in Chapter 9.

Notably, in regions showing linear effects (ROI defined from the second-level *linear* effects maps), we find again present the quadratic effects. In addition, when the quadratic

parametric modulations were not orthogonalized on the linear parametric modulations, we found only a quadratic effect and no linear effect in vmPFC. In MCC, both linear and quadratic effects were maintained.

Negative quadratic effects. We found in lateral PFC a negative quadratic effect only for the belief system (left: MNI peak coordinates: [-36, 59, -5], T = 4.93, right: MNI peak coordinates: [36, 56, -8], T = 6.82) (Figure 8.27), covering a large region from BA8/BA46 up to frontopolar cortex (BA10). However, our model fitting showed that the belief system contributed more in the mixture. Indeed, the weight of the affective values system contribution in the mixture was $\omega = 0.25$ on average, with $\omega < 0.1$ for more than a third of subjects. Therefore, we propose that this negative quadratic effect was more likely related to global decision values, supporting a role for lateral PFC in performing action selection.

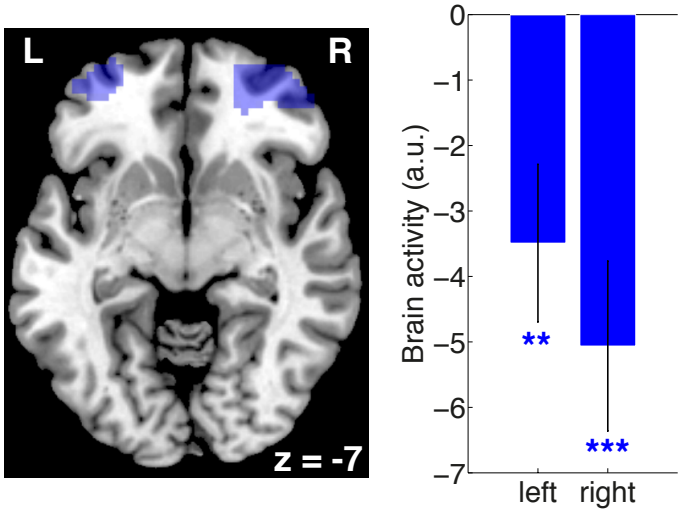


FIGURE 8.27: Involvement of lateral PFC in action selection.
Left panel: axial slice of parametric brain activations negatively correlating with relative chosen belief^2 (blue) thresholded at $p < 0.005$ (voxel-wise, uncorrected). z is brain slice MNI coordinate. Right panel: Effect sizes for relative chosen belief^2 for left and right lateral PFC clusters, averaged over voxels from a sphere of radius 13 mm centered on the activation peak. a.u. arbitrary units. Error bars correspond to s.e.m. across subjects (N = 21). $^{**}p < 0.01$, $^{***}p < 0.005$.

Controlling for reaction times. We checked that the quadratic effects were maintained despite the inclusion of an additional parametric modulation coding for reaction times. All brain activations were maintained, in particular the double-dissociation between beliefs and affective values regarding choice-independent effects in vmPFC and MCC respectively. The only difference was that the clusters, at the same threshold, contained slightly less voxels (54 voxels instead of 99 for the MCC cluster, 797 voxels instead of 862 for the vmPFC cluster). Reaction times monotonically decreased with

both belief chosen and decision value chosen (Figure 8.28). The higher the belief, the faster the subject decided. Yet, reaction times are a good proxy of choice difficulty. Importantly, reaction times did not scale quadratically with belief chosen. Therefore, we could rule out an interpretation of the quadratic effects simply in terms of choice difficulty. Figure 8.28 nicely illustrates the impact of the sampling method for binning the data, as further explained in the Methods section. There were two possibilities to build one-dimensional bins. On the one hand, bins could be constructed with fixed boundaries of the same size, resulting in a different number of events per bin (top panels, Figure 8.28). On the other hand, bins could be constructed with quantiles including the same number of events per bin, but resulting in bins with variable boundaries (bottom panels, Figure 8.28).

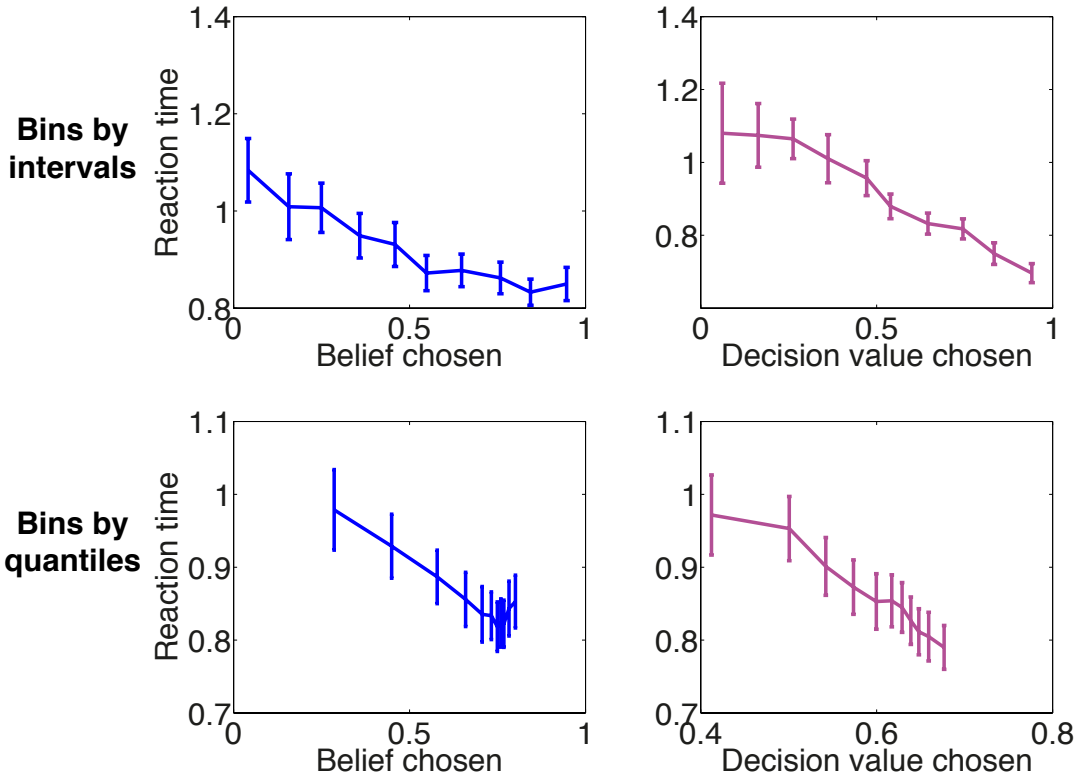


FIGURE 8.28: Reaction times monotonically decreased as a function of belief chosen (left panels) and decision value chosen (right panels). Two sampling methods for building the bins are illustrated. Bins could be either constructed using intervals with fixed boundaries but a variable number of trials per bin (top panels) or with variable boundaries but the same number of trials per bin (bottom panels).

Controlling for stay/switch trials. We checked that the quadratic effects were maintained despite the inclusion of an additional parametric modulation coding for stay/switch trials, according to whether the subjects picked the same shape as in the previous trial or not. All brain activations were maintained, in particular the double-dissociation between beliefs and affective values regarding choice-independent effects in

vmPFC and MCC respectively. However, the clusters, at the same threshold, contained less voxels (82 voxels instead of 99 for the MCC cluster, 64 voxels instead of 862 for the vmPFC cluster). Here, it seems that there was shared variance between the belief and the stay/switch parametric modulation. In fact, a similar vmPFC cluster was found to correlate with "stay" trials. This result is consistent with the interpretation that the belief system monitors the current state. In addition, the vmPFC cluster previously found to correlate linearly positively with the relative chosen affective value was almost absent (28 voxels left), maybe also in relation with stay trials.

Controlling for both reaction times and stay/switch trials. We checked that the quadratic effects were maintained despite the inclusion of two additional parametric modulations. One coded for reaction times and the other coded for stay/switch trials, according to whether the subjects picked the same option as in the previous trial or not. Critically, the presence of a double-dissociation between beliefs and affective values regarding choice-independent effects was maintained. The vmPFC and MCC clusters, however, at the same threshold, contained less voxels (47 voxels instead of 99 for the MCC cluster, 41 voxels instead of 862 for the vmPFC cluster). Otherwise, the same modifications were observed as when controlling for stay/switch trials only. Obviously, activations generally become smaller with the growing number of parametric modulations.

Notably, all reported activations were maintained despite methodological choices. When β coefficients (effect sizes) were extracted on a small sphere centered on the peak voxel of the cluster instead of averaged over all voxels of the cluster, the obtained statistics were very similar. Moreover, it should be emphasized that the β extraction of effect sizes was done in a ROI defined from an independent analysis, using a leave-one out procedure. Lastly, the role of the full variance analysis (sometimes called "unique variance analysis") is to be highlighted. All GLM were run in full variance analysis, meaning that observed variance for each parametric modulator was specific. The global shape of the activations as presented with the bins analyses was dependent on the sampling procedure, as explained in detail in the Methods section. However, we checked that the activations shape was qualitatively similar whatever the method used for binning the data.

Remarkably, when we replaced our quadratic regressors by regressors coding absolute values instead of squares, we reproduced all the effects. So it could be a U-shaped effect as well as a V-shaped effect. Therefore, we do not push the interpretation of quadratic effects beyond a simple effect of absolute value, meaning, unsigned by choice.

To sum up, linear effects (signed by choice) more probably corresponded to choice-dependent, post-choice signals, whereas quadratic effects (unsigned by choice) more probably corresponded to choice-independent, pre-choice signals.

Conclusion. Taken together, these results suggest the following architecture for the integration of beliefs and affective values in human decision-making (Figure 8.29). Before decision, vmPFC and MCC separately encode beliefs and affective values respectively, as supported by the double-dissociation between vmPFC and MCC regarding choice-independent signals. Lateral PFC combines both signals to decide. In return, lateral PFC feeds back choice information to these medial regions, presumably for updating these value signals according to action outcomes, hence choice-dependent representations in both vmPFC and MCC.

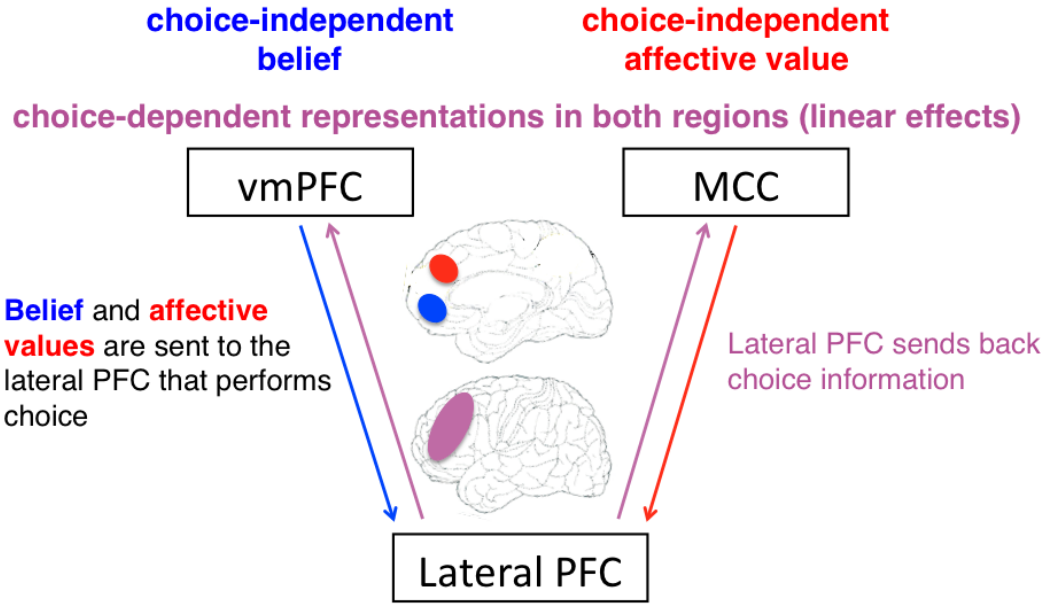


FIGURE 8.29: Interaction between lateral and medial PFC in decision-making. Before decision, vmPFC and MCC separately encode representations for the belief system and the affective values system respectively. Both components are then transferred to the lateral PFC which combine them to make a decision. After choice, lateral PFC sends back choice information to the medial regions, which in turn updates representations within the two systems.

8.3.3 Replication of results in an independent analysis

To visualize linear and quadratic effects in the functional data, we sorted trials in a 2D-grid according to two dimensions: belief and affective value. Details about the bins construction are provided in the Methods section. Essentially, mean brain activity was estimated in each bin. Critically, in MCC and vmPFC ROIs defined from parametric

maps of positive quadratic effects in GLM2, we observed linear and quadratic effects consistent with the statistics obtained for GLM2 second-level parametric maps. (Figure 8.30). Positive linear effects were found in vmPFC for both belief and affective value. Negative linear effects were found in MCC for both belief and affective value. A positive quadratic effect was specifically found in vmPFC for belief, and specifically found in MCC for affective value.

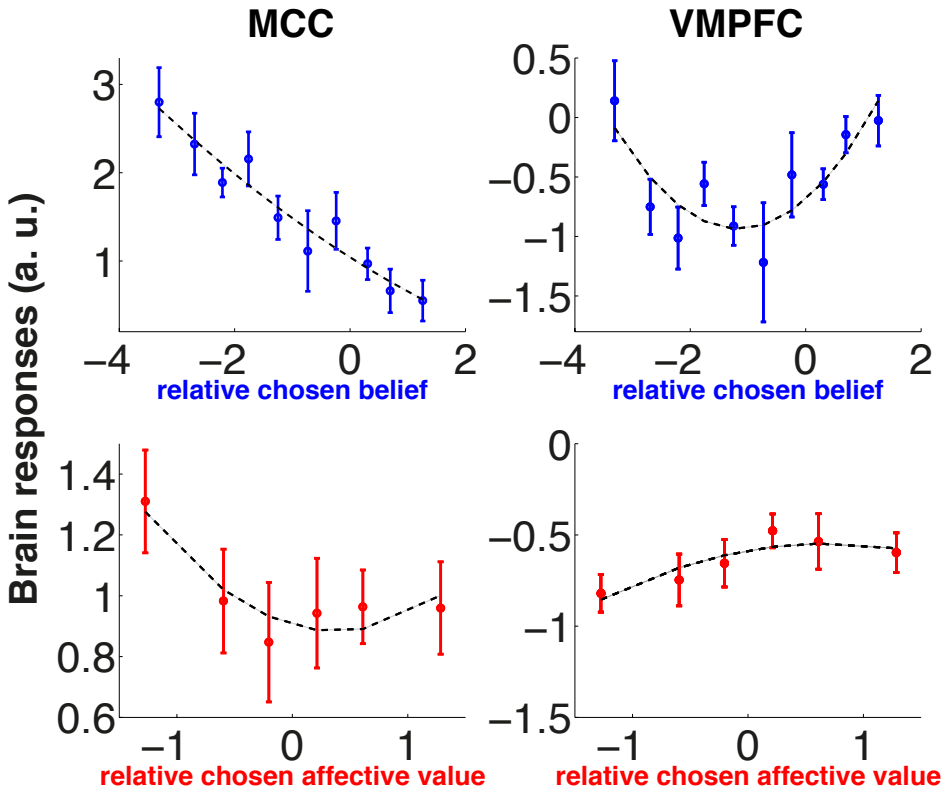


FIGURE 8.30: Bins analysis in ROIs defined from positive quadratic effects parametric map. In an independent analysis, we reproduced choice-dependent and choice-independent representations of both beliefs and affective values in our two main regions of interest, MCC (left) and vmPFC (right). Dashed lines show the best polynomial fit of degree 2. Error bars represent s.e.m. across bins (all subjects pooled).

Importantly, when we reproduced this analysis using ROIs defined from parametric maps associated with linear effects instead of quadratic effects, we observe very similar shapes of the brain responses (Figure 8.31). Thus, the quadratic effects were present in the signal even in ROIs isolated using linear effects parametric maps.

8.3.4 GLM3: Further dissociation within each system

Finally, we tested whether there were distinct neural correlates for the four quantities involved in the mixed model to compute decision values (Figure 8.11). BOLD signal

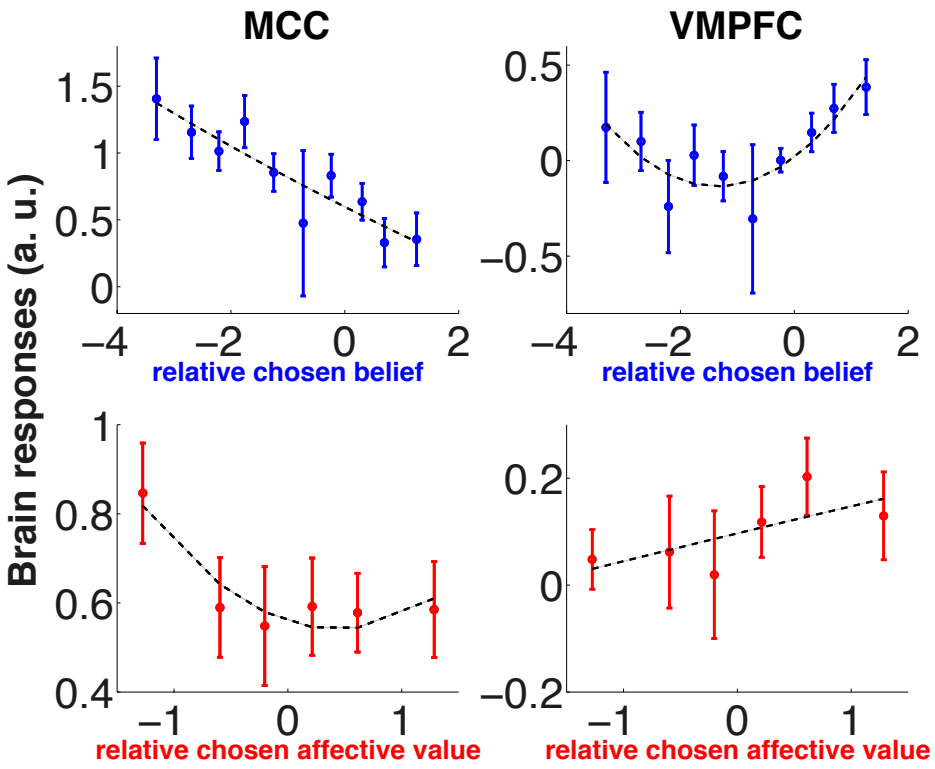


FIGURE 8.31: Bins analysis in ROIs defined from linear effects parametric maps. In an independent analysis, we reproduced choice-dependent and choice-independent representations of both beliefs and affective values in our two main regions of interest, MCC (left) and vmPFC (right). Dashed lines show the best polynomial fit of degree 2. Error bars represent s.e.m. across bins (all subjects pooled).

was regressed against linear and quadratic effects of (1) Beliefs, (2) Informational values of proposed rewards (likelihood), (3) Reinforcement values and (4) Affective value of proposed rewards, in unique variance, meaning that the observed activations in the following second-level maps was again truly selective of each variable.

8.3.4.1 Dissociation within the affective values system: Reinforcement values (historical) vs. Affective values of proposed rewards (current trial)

The affective values system in our mixed model was composed of two variables: *Reinforcement values*, which are the historical variables learnt by reinforcement learning across trials, and *Affective values of proposed rewards*, which are the potential rewards to gain displayed before each choice.

Regarding choice-independent effects, we found a positive quadratic effect of Reinforcement values in lateral orbitofrontal cortex (MNI peak coordinates: [39, 38, -17], T = 3.47) (Figure 8.32).

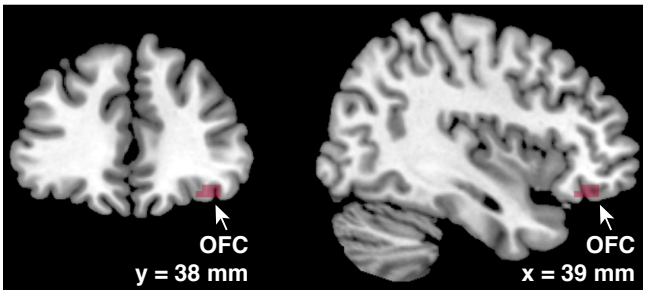


FIGURE 8.32: Within the affective values system, lateral OFC encoded choice-independent reinforcement values. Coronal and sagittal slices of parametric brain activations positively correlating with relative chosen reinforcement value^2 thresholded at $p < 0.005$ (voxel-wise, uncorrected). Coordinates of brain slices correspond to the activation peak (MNI space).

8.3.4.2 Dissociation within the Bayesian system: Prior belief (historical) vs. Informational values (current trial)

Similarly, the Bayesian system in our mixed model was composed of two variables: *Prior beliefs*, which are the historical variables learnt from the past, and *Informational values of proposed rewards (likelihood)*, which are extracted from proposed rewards displayed before each choice.

Regarding choice-independent effects, we found a positive quadratic effect of Informational values of proposed rewards in precentral gyrus (MNI peak coordinates: [39, -16, 40], T = 3.76) (Figure 8.33).

8.3.5 Activations at feedback time

At feedback time, the dissociation between the two value systems was less clear. Here, we report the main findings at the feedback reception onset, with a focus on frontal lobes activity.

We observed that vmPFC activity correlated positively with the reward received. We also found a positive effect of reward received in bilateral putamen and pallidum, as observed in previous studies of value-based learning and decision-making. In addition, dACC correlated negatively with reward received.

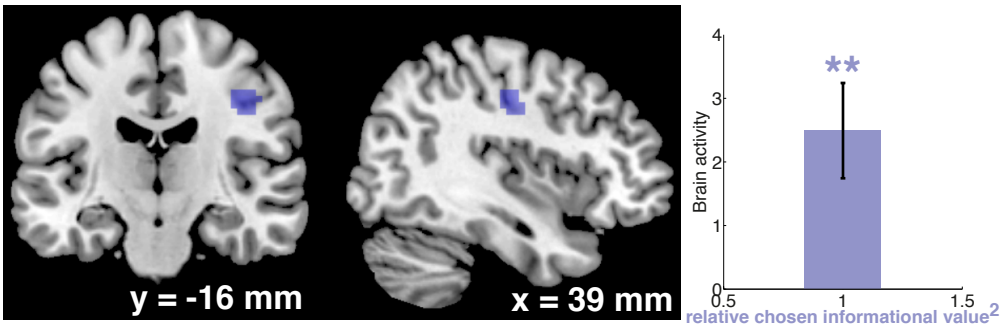


FIGURE 8.33: Brain implementation of informational values from proposed rewards extracted at decision time in pre-central gyrus.
Left panel: coronal and sagittal slices of parametric brain activations positively correlating with relative chosen informational value^2 thresholded at $p < 0.005$ (voxel-wise, uncorrected). Coordinates of brain slices correspond to the activation peak (MNI space). Right panel: Effect size, averaged over voxels from a sphere of radius 13 mm centered on the activation peak. a.u. arbitrary units. Error bars correspond to s.e.m. across subjects (N = 21). **$p < 0.01$.

We observed a large vmPFC cluster correlating positively with the belief chosen, as well as a bilateral temporal activation. ACC, MCC and frontopolar cortex (bilateral) correlated negatively with belief chosen, as well as bilateral insula. Dorsolateral PFC (bilateral) also correlated negatively with belief chosen.

Chapter 9

General Discussion

In this PhD work, we addressed the question of how beliefs and affective values conveyed by rewards contribute to decision-making.

We presented several behavioral experiments dissociating these two signals, in the form of probabilistic reversal-learning tasks involving stochastic and changing reward structures. We built a model establishing the functional and computational grounds of such a dissociation. It combines two parallel systems: (1) reinforcement learning, modulating affective values, and (2) Bayesian inference, monitoring beliefs. The model better accounted for behavior than many alternative models.

We then investigated whether beliefs and affective values have distinct neural bases using fMRI. We showed that VMPFC and MCC activity correlated with both choice-dependent beliefs and affective values. However, we found a double dissociation regarding choice-independent variables. VMPFC encoded choice-independent beliefs, while MCC encoded choice-independent affective values. Additionally, activity in LPFC increased when decision values (i.e. mixture of beliefs and affective values) got closer to each other and action selection became more difficult.

Taken together, these results suggest that before decision, VMPFC and MCC encode beliefs and affective values respectively. LPFC combines both signals to make a decision, then feeds back choice information to these medial regions, presumably for updating these two signals according to action outcomes.

We will first discuss the functional interpretation of our computational mixed model and its limits. We will then come back to the imaging results and discuss especially the role of ventromedial prefrontal cortex, as well as the meaning of the quadratic effects that we found.

9.1 Modeling

9.1.1 Distortions

9.1.1.1 Differences with prospect theory

Behavioral economics investigate how subjects combine probabilities and values. In a typical experiment (Kahneman, 1984 [170]), subjects undergo a choice between two gambles, for which they know the probability and gain at stake. In such a situation, the optimal choice would be to compute an expected value (probability times reward) for each gamble, and choose the gamble of highest expected value. Usual prospect theory results by Kahneman and Tversky showed that human subjects' choices generally depart from optimality. In this framework, they typically tend to overweight small probabilities, even if such outcomes are rare, and underweight large probabilities, resulting in a S-inverse shaped distortion function (Kahneman and Tversky, 1979 [98]).

However, the level of description here is different from that of neuroscience. Behavioral economics and psychology describe the existing biases and departures from optimality in economic decision-making, but do not investigate the underlying brain mechanisms producing such outputs.

At first sight, the fact that we observed distortions on probabilities that were sigmoidal could seem puzzling, in contradiction with Kahneman and Tversky primary results. However, there were two major differences between our protocol and the prospect theory framework. First, in our experiment, probabilities, in the form of beliefs about how actions map onto reward contingencies, were implicit. They had to be estimated from experience. Subjects did not have an explicit knowledge of probabilities. On the contrary, in characteristic prospect theory experiments, subjects typically had to choose between two gambles with an explicit knowledge of probabilities and reward values at stake (Kahneman and Tversky, 1979, 1974 [98] [171]). Consequently, a second difference was related to the uncertainty level. While in prospect theory gambles, subjects had a perfect knowledge of the probabilities at stake, in our protocol subjects had to estimate probabilities directly from experience, by trial-by-trial sampling. Therefore, subjects' probability estimates in our task were accompanied by some uncertainty, with some noise in the underlying brain representations, since neuronal computations are not perfect (Beck et al., 2012 [184]).

The fact that subjects tend to overweight small probabilities when these are directly experienced has actually been documented before (Hertwig et al., 2009 [185]). In this

study, Hertzig and colleagues noticed that choices from described probabilities considerably differed from choices from experienced probabilities. In particular, they found an underweighting of small probabilities when decisions were made from experienced probabilities, consistent with the sigmoid-like distortion that we found.

In another study, Tobler and colleagues (Tobler et al., 2008 [186]) even reported various distorted perception of experienced probabilities depending on the brain region. More precisely, they found in dorsolateral PFC an overweighting of small probabilities and an underweighting of large probabilities, while ventral PFC regions displayed the opposite representations. However, their paradigm involved valuation without choice; the absence of decision might have led to different results.

Yet, even if distortions do not explain what happens in the brain, they remain a useful tool at the behavioral level. In a clinical context, distortions have been used to study the rewards perception in pathological gambling, and more generally, in addictions. Patients with addiction display a higher discount rate and a stronger preference for immediate rewards (Michalczuk et al., 2011 [187]). Distortions are also a tool to study the probabilities perception in the context of decision-making under risk (Sharp et al., 2012 [105]).

9.1.1.2 Sub-optimality and efficient coding

Even though distorting probability representations might seem suboptimal, Summerfield and colleagues proposed that such an underweighting of small probabilities when decisions are made from experience, related to extreme or rare events, could actually be understood as optimal in terms of *efficient coding* (Summerfield et al., 2015 [188]). Down-weighting of low probability events would allow more robust coding. In terms of computational resources, it is not efficient to dedicate energy to encode very rare events that will almost never be encountered. Efficient coding thus explains why rare events are not represented: it is not optimized in terms of computational resources to represent rare events. In other words, efficient coding allows optimal and appropriate allocation of cognitive resources. It allows more adequate coding of the information relevant for decision, given the statistics of the current environment. Further mathematical details about how specific encoding of extreme events can actually be seen as rational were provided by Lieder et al., 2014 [189].

Similarly, coding expected values i.e. a product is more computationally demanding in terms of resources than coding a sum. According to the mixed model, probabilities about states (belief) and rewards (affective values) are *linearly* combined. Indeed, the mixed model (linear) was a significantly better fit than the Bayesian model alone or the

distortions model (both multiplicative). In other words, we found that a linear combination better explained subjects' behavior than a multiplicative combination. Our mixed model thus constitutes a simplification of the computation, as compared to a more costly computation of expected values, thus optimizing resources. It suggests that subjects tended to simplify the task, as compared to an optimal agent. We cannot formally exclude that they simplify the task due to performing hundreds of trials, but still, this simplification has an interest, since computing a sum demands less computational resources than computing a product i.e. expected value.

9.1.2 Predominance of the belief system

Our mixed model fit revealed that the belief system, which is a monitoring system, is predominant for guiding choice. Choice can be biased by affective values when the affective values significantly depart from the belief.

The distortions model fit revealed that probabilities were distorted in a sigmoid fashion. Subjects tended to underweight lower probabilities while overweighting higher probabilities. It means that they had a propensity to accentuate their representations in a binary manner: correct vs. incorrect action, with little parametric modulation. It means that to maintain expected values, expected values had to be distorted. We can abstract from distortions with the mixed model, more simple. The shape of the fitted distortions was actually consistent with the mixed model fits, which predicted a larger contribution of beliefs than affective values in the mixture for decision.

Therefore, we do not formally exclude the distortions model but we think of it as a different level of description. Regardless of the brain, the distortions model consists of a good description of behavior. However, the mixed model is a simpler model. Moreover, it provides a psychological origin for the distortions. In other words, it gives a mechanistic explanation of why distortions arise: two concurrent systems act in parallel to guide decision-making.

Importantly, according to both the mixed model and the distortions model, the belief system, based on hidden states, consistently remains predominant in the decision. This constant predominance of beliefs in the mixture suggests that the system is intrinsically built this way. It allows to base decisions on stable representations, continuing across trials. Relying mainly on a belief allows more stability, as compared to recalculating expected values at each trial.

Lastly, the mixed model makes a decision based on a mixture of variables from the two systems at each trial. Although this is a good *average* description, it is possible

though that in certain trials, choice was based on belief, whereas in other trials, choice was based on affective values. There could be an alternation between the two systems over the course of trials, rather than a similar relative contribution of both systems in a mixture occurring at each individual trial. For example, we could hypothesize that after a reversal, subjects would rely more on the affective values system. In contrast, at the plateau of a behavioral episode, meaning a period of relative stability, subjects would rely on the Bayesian system. In a perceptual decision-making task, Summerfield and colleagues found such an alternation of two systems driving decisions (working memory vs. Bayesian) according to the trial position relative to the reversal (Summerfield et al., 2011 [190]). A finer refinement of our model over time could allow to test these possibilities.

9.1.3 Interaction between the belief system and the affective value system: a hierarchy?

We believe that our best-fitting mixed model, which consists of a mixture of two systems for decision-making, is a reasonable approximation of the real variables calculated in the brain. Critically, we replicated the mixed model supremacy to explain behavior across two different protocols, and with three experimental variants for each protocol (about 25 subjects each). However, there are two possible interpretations for this mixture, regarding the nature of the interaction between the two systems, the belief system and the affective values system. A first interpretation is that the belief and the affective value systems are combined into a single decision value, via a summation. An alternative interpretation is that there is a hierarchy behind the mixture, with two successive decisions. However, our study alone does not allow to disentangle these two possibilities. How do the Bayesian system and the affective values system actually interact?

An interesting avenue for future research would be to investigate whether there is a hierarchy between the two systems in the mixed model. Hierarchical brain systems have been described previously in the literature. For example, a hierarchy in information processing for cognitive control was described within human lateral prefrontal cortex (Koechlin et al., 2003 [36]). We could hypothesize that the affective values system (RL) would act as a default, insuring learning in a great number of situations (Doll et al., 2012 [191]). The beliefs system (Bayesian) would be recruited only in necessary situations, for example when the environment is more uncertain and/or more complex, subsequently requiring more cognitive control.

The alternative hypothesis in terms of hierarchy would be that the beliefs system would be the predominant one. The observed predominance of the beliefs in the mixture

supports the latter interpretation. In that case, while being slightly biased by the affective values system in decision, subjects would mainly rely on prefrontal function to find which is the more appropriate action to select in a given situation, evaluating to what extent the current behavior matches external contingencies. In situations with a large uncertainty or a high volatility, the prefrontal system would take control to infer the most appropriate action to select. Nevertheless, this might stand only until a certain level of complexity. In complex and uncertain situations, there is a limit in terms of cognitive load subjects are able to handle (Oberauer, 2002 [44]; Collins and Koechlin, 2012 [160]). In too complex/cognitively demanding tasks, subjects might go back to simpler strategies, such as pure model-free reinforcement learning.

9.1.4 Difference with model-based/model-free reinforcement learning

In model-free RL, subjects learn stimulus/action pairs by directly experiencing them in the external environment. By contrast, in situations in which it is beneficial to include a knowledge of the task's contingencies structure, subjects were able to develop internal model-based representations, thus improving learning efficiency (Hampton et al., 2006 [56]; Behrens et al., 2007 [162]). In this framework, arbitration between model-based and model-free RL systems was based on the computation of each system's uncertainty (Daw et al., 2005 [110]). As discussed in the introduction, in a two-stage probabilistic decision-making task, subjects' behavior appears intermediate between model-based and model-free RL systems (Dezfouli and Balleine, 2013 [192]).

At first sight, this distinction between model-based and model-free RL resembles our mixed model. Indeed, a parallel could be drawn here between our affective values system and model-free learning, and between our beliefs system and model-based learning. However, there is a fundamental difference between our mixed model account and the model-based/model-free RL account. Specifically, in model-based RL, the model is used to calculate expected rewards. Expected rewards are computed based on the model's states representation. Model-based expected rewards are then mixed with the expected rewards from the model-free system. In contrast, the model's states representation in our mixed model is used to calculate a degree of belief about the mapping between states and outcomes. The belief in our mixed model is used for inferential reasoning, not for computing expected rewards.

In addition, a limit of the model-based/model-free approach is that as soon as the number of states and the number of stimulus/action pairs increases, model-based reinforcement learning becomes intractable.

9.1.5 Prefrontal cortex: a not yet optimized system?

We consistently observed, across the various behavioral experiments presented here, that subjects are sub-optimal. They behave irrationally as compared to a statistically optimal agent. More precisely, they are suboptimal in the sense that they do not maximize expected reward. In other words, their choices do not maximize expected value. This sub-optimality in value-based decision-making has been reported many times (Payzan-LeNestour and Bossaerts, 2011 [144]). It could be due to the fact that the prefrontal system is phylogenetically very recent (Teffer and Semendeferi, 2012 [193]). Prefrontal cortex functional architecture is shaped by very recent evolution, in complex social environments, possibly explaining why it is less well optimized as compared to other systems. For example, the visual system functioning is well explained by optimal Bayesian models of perception (e.g. Kersten and Yuille, 2003 [194]). The visual system is much older in phylogenetic evolution, which can explain the fact it is so well optimized. By contrast, the prefrontal system is much more recent.

In addition, the central executive system faces very diverse problems, of high complexity. There is a huge variety of tasks that the prefrontal cortex deals with: planning, social interactions, task-switching, arbitrating for cognitive resources allocation, etc. It is recruited when peripheral systems are no longer efficient. Facing such a diversity of tasks, it is thus impossible for the prefrontal system to be optimal for this, since it would be computationally intractable, too demanding in terms of cognitive resources. By contrast, the visual system is well adapted and optimized for one particular task only (e.g. visual detection).

These two arguments might explain why subjects behave sub-optimally in decision-making tasks.

9.1.6 Beliefs, affective values, and stability of representations

We provided evidence supporting the idea that action outcomes convey two types of value signals: affective values and belief values. We revealed that choice-independent representations of beliefs and affective values are encoded in vmPFC and MCC respectively.

A possible alternative interpretation would be that vmPFC would encode a relatively stable variable over the course of a behavioral episode (belief), whereas MCC would encode a less steady variable, changing more rapidly from trial to trial (affective value). This alternative interpretation is in line with recent results by Tsetsos and colleagues (Tsetsos et al., 2014 [195]). The authors showed representations in vmPFC (rostral)

when a decision was deferred to later, i.e. a stable, long-term maintenance of values relevant for decision. In contrast, the dACC, equivalent to our MCC cluster, was activated when subjects committed to a choice in the current trial, i.e. variables relevant to decision in the short term. Thus, in that framework, representations in MCC were less stable.

To examine this alternative "stable/less stable" interpretation as compared to our "belief/affective value" interpretation, we tested whether in our experiment, affective values varied more rapidly from trial to trial than beliefs. For each subject, we calculated the absolute difference between the belief at trial t and at trial $t+1$, and compared it to the absolute difference between the affective value at t and at $t+1$. We used the absolute difference because we were not interested in the direction but in the amount of variation from trial to trial. We showed that, from trial to trial, affective values estimates were marginally less steady than beliefs estimates (paired t-test across subjects: $p = 0.045$). In fact, in our reversal-learning task, the belief associated with each shape was quite constant between two reversals. After a reversal, subjects stabilized their belief about which shape is the most frequently rewarded one, until the next reversal.

In addition, when we include an additional parametric modulation coding for stay/switch trials, the positive quadratic effect of affective values in MCC remained present. Yet, coding a switch is coding a short-term event. This rules out the possibility that MCC would simply encode the short-term evidence in favor of a switch. Rather, this result supports our interpretation of MCC encoding a choice-independent affective value representation.

However, the notion of belief in our model refers to a belief about how actions map onto outcome contingencies. It relates to internal mental states. Therefore, perhaps subjects build a belief about how their actions match the current external contingencies only in relatively stable environments, for the long run. Indeed, the notion of belief relates to a higher-order structure based on hidden states. Beliefs might not be involved in overly variable and uncertain environments. Indeed, if contingencies are permanently changing, it is not relevant to build a belief. More broadly, it could be that the notion of belief, based on abstract hidden states, would be intrinsically more stable.

9.2 Imaging results: understanding the role of vmPFC and MCC

Using fMRI, we found choice-dependent representations of beliefs and affective values in both vmPFC and MCC. We revealed that choice-independent representations of beliefs

and affective values were dissociated, encoded in vmPFC and MCC respectively.

9.2.1 vmPFC and reliability signals

In our team, during the last few years, a mathematical theory of task-sets monitoring in open-ended and changing environments has been developed, based on reliability signals inferred from past outcomes ("PROBE model", Collins and Koechlin, 2012 [160]). The reliability signal associated with the current behavioral task-set was found to be encoded in vmPFC (Donoso et al., 2014 [94]).

The concept of belief in the present model is very similar to the notion of reliability in the PROBE model framework [160] [94]. Essentially, it relates to the subjects' ability to infer how well their behavior matches current external contingencies, i.e., how good their representations are predictive. These belief/reliability signals are then used to monitor behavior in relation to internal mental states, and to infer when external contingencies have changed, in order to subsequently switch behavior if necessary. This ability to flexibly adapt to the current environmental situation is at the core of prefrontal function.

In line with our imaging results, another study reported that vmPFC activity was rather consistent with an abstract state-based inferential model (beliefs) than with a reinforcement learning model (affective values) (Hampton et al., 2006 [56]). In a probabilistic reversal-learning task under fMRI, subjects had to choose between two stimuli, for which the reward contingencies were anti-correlated. The task was quite hard because soon after subjects identified the "good" option to choose, the two options reversed. After a negative outcome, or a series of negative outcomes, subjects more often switched options. Hampton and colleagues reasoned that when subjects decide to switch, a reinforcement learning model and a state-based model would make different predictions. On the one hand, if subjects use a reinforcement learning model, the value of the newly chosen option should be low, because it was low the last time the subject chose it and subsequently abandoned it. On the other hand, if subjects use a state-based model, the value of the newly chosen option should be high, because the subject inferred the underlying task structure. He inferred that the two options are anti-correlated (if one is low, the other is high), so when he abandoned a low-valued option, he knows that the value of the newly chosen option should be high. The authors observed that qualitatively, BOLD signal in vmPFC was rather consistent with a state-based inferential model than with a reinforcement learning model (Figure 9.1).

Lastly, we provided here fMRI data, which is only correlative. With model-based fMRI, we search for regions in which brain activity correlates with specific variables. To evidence the vmPFC causal responsibility in belief modulation, we could use transcranial

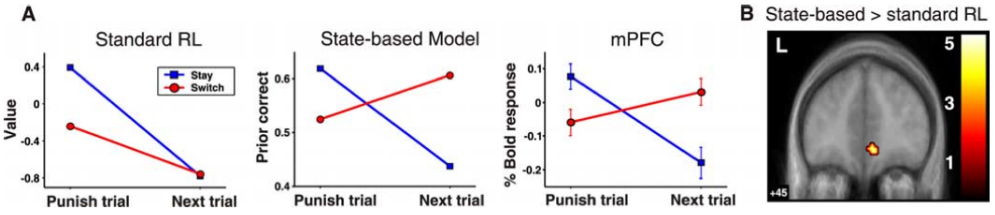


FIGURE 9.1: In switch trials (red lines), fMRI activity in vmPFC was rather consistent with a state-based model (belief values) than with a reinforcement learning model (affective values) (reproduced from Hampton et al., 2006).

magnetic stimulation (TMS). TMS consists in externally applying magnetic impulses. Consequently, the activity of neurons under the magnetic field is modified. Another possibility would be to examine the behavior of patients with a focal lesion in vmPFC. Nonetheless, other methodological limits arise with TMS or lesion studies. In particular, the connectivity between the target brain region and other areas is also modified.

9.2.2 vmPFC and the default mode network

We observed that the vmPFC represented choice-independent belief. vmPFC is a central node of the default mode network and is involved in the monitoring of internal mental states and mentalizing (Esposito et al., 2006 [196]; Fransson, 2006 [197]; Gusnard et al., 2001 [198]). The default mode network is thought to mediate the attentional engagement towards internally or externally oriented tasks. This view is consistent with our finding that vmPFC monitors a belief about how actions map onto current external outcome contingencies, in line with reports that anterior PFC arbitrates between "in" and "out" attentional modes (Burgess et al., 2007 [88]).

9.2.3 vmPFC and the notion of value

We have seen in chapter 3 that vmPFC encodes subjective affective values, as supported by a large number of empirical studies in both animals and humans. Importantly, we replicated in our dataset the classic effect of chosen value encoding in vmPFC (Plassmann et al., 2007 [53], Chib et al., 2009 [52]). This was found in a number of neuroeconomics studies, using primary rewards as well as secondary rewards (reviewed in Clithero and Rangel, 2013 [199] and in Sescousse et al., 2013 [179]). In these studies, it is possible that the positive linear effect reported for chosen values was masking an effect that was actually quadratic. Indeed, to our knowledge, both linear and quadratic effects were not modelled within the same analysis. With our full variance analysis, we

were able to separate activations selective to linear and to quadratic effects respectively, and we showed that vmPFC activity was rather consistent with a quadratic effect of the belief.

Moreover, quadratic effects associated with value have been reported before (Padoa-Schioppa and Assad, 2006 [48]).

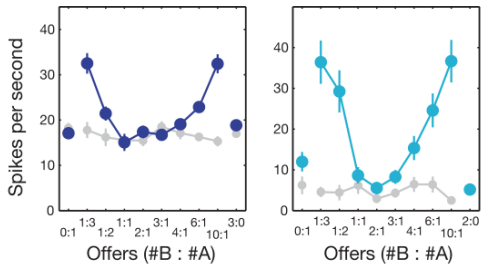


FIGURE 9.2: Pattern of neurons encoding the juice chosen value (reproduced from Padoa-Schioppa and Assad, 2006).

Certain neurons in OFC (area 13m in monkeys) encode the chosen value of juice displayed a typical quadratic pattern (Figure 9.2). These neurons fired more when the chosen and unchosen juice values were far from each other and less when they were close, independently of visuomotor contingencies. In this study, the concept of value corresponds to the notion of affective value in our terminology, with juice acting as a primary reward. So, the OFC neurons quadratic activity pattern in Figure 9.2 is very likely to correspond to that of the lateral OFC cluster that we found for reinforcement affective values (GLM3). Although Padoa-Schioppa and Assad's paradigm did not involve learning, there could have been a reinforcement effect due to the high number of trials that the monkeys performed. Consistently with previous studies probing the role of OFC, the lateral OFC cluster that we observed for the quadratic expansion of reinforcement affective values could have a role in imagining/inferring future outcome, as reinforcement affective values corresponds to the average expected value anticipated.

In addition, in certain previous studies, the concept of value was sometimes not precisely defined (O' Doherty, 2014 [54]). Outcome value has been confounded for example with outcome identity (Klein-Flügge et al., 2013 [200]), outcome sensory properties, outcome predictive information, outcome informational value (Jessup and O'Doherty, 2014 [201]), outcome saliency in terms of attentional effects (Kahnt et al., 2014 [202]). This could explain why quadratic effects have not been emphasized before, probably being hidden behind linear effects due to asymmetric sampling.

Salience account. Lastly, we can rule out an interpretation in terms of salience for the quadratic effects reported here, under the following argumentation. The concept of

salience or saliency refers to stimuli drawing attentional resources and able to enhance perceptual, cognitive or emotional processing of these stimuli. It also refers to the stimuli motivational properties for action (Kahnt et al., 2014 [202]). One could interpret the quadratic effect of value reported here as a salience effect. Indeed, very high valued outcomes or very low valued outcomes are likely to engage more attentional resources. However, the quadratic effect reported here is a function of *value chosen minus unchosen*. So, if participants had to choose between 10 Euros and 10 Euros as proposed rewards, the trial overall salience would be high. In contrast, if participants had to choose between 2 Euros and 2 Euros as proposed rewards, the scene general salience would be low. But in both cases, the quantity *value chosen minus unchosen* would be the same. Therefore, the quadratic effect of value and saliency were unrelated in our paradigm, which allows us to rule out an interpretation in terms of salience.

9.2.4 vmPFC and the notion of confidence

We found a quadratic U-shaped effect of belief in vmPFC, which means that vmPFC activity increased at the extreme of the belief scale, whereas vmPFC activity was the lowest in the middle. We argued that this effect might rather reflect unsigned pre-choice preferences, irrespective of choice, rather than a post-choice confidence signal. Indeed, we expected confidence to be a post-choice global signal, not dissociated. However, confidence signals in vmPFC have been reported in previous studies (De Martino et al., 2012 [203]). In a choice task between two food items followed by subjective confidence reports in having made the best decision, De Martino and colleagues viewed confidence as the absolute difference between two accumulators at decision time [203].

Whereas the idea that confidence is a second-order construct is supported by several studies, the exact underlying computation of confidence remains unclear. For instance, confidence has been viewed as a read-out of the noise in the decision process (De Martino et al., 2012 [203]), or including action information (Fleming et al., 2015 [204]), or automatically arising (Lebreton et al., 2015 [205]). In a recent study, Lebreton and colleagues revealed quadratic U-shaped effects of confidence in vmPFC, in a series of tasks involving subjective ratings (Lebreton et al., 2015 [205]). More precisely, they elegantly revealed that following ratings of pleasantness, probability, age or desirability of stimuli, a U-shaped signal encoding confidence was automatically found in vmPFC (Figure 9.3). The confidence signal in vmPFC was present even in the absence of explicit confidence reports and even if it was not required for the current task. It could not be explained by either saliency or valuation accounts.

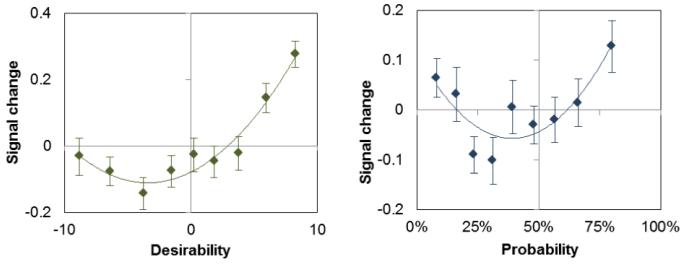


FIGURE 9.3: Quadratic effects of confidence observed in vmPFC elicited with a desirability rating task (left panel) and a probability rating task (right panel) (reproduced from Lebreton et al., 2015).

Our results are in line with this study. Yet, the authors did not dissociate the concept of belief and the concept of affective value. Indeed, they considered that being accurate (belief) in a judgment is valuable (affective value) in itself. Another difference between our paradigm and theirs was the presence/absence of choice, even if a rating could be considered as a choice, mapping internal subjective preferences to an external scale for report. However, specifically in the case of pleasantness and desirability ratings, which both include an aspect of affective value as we conceptualized it in our work, the conjunction with confidence signal revealed that the vmPFC cluster was accompanied by a more dorsal cluster, which fits very well with the MCC U-shaped cluster observed for affective values in our study.

Therefore, it could be that the quadratic signals we observed might not be confidence per se. Rather, it might be a signal contributing upstream to the construction of a confidence measure. Metacognition refers to humans' ability to evaluate and monitor their own behavior, linking objective performance to subjective confidence (Fleming and Dolan, 2012 [96]). The U-shaped signals we observed could thus support the brain implementation of metacognitive processes.

9.2.5 MCC and the affective values representation

Our data revealed a positive quadratic effect of affective values in the midcingulate cortex. MCC activity was higher when chosen and unchosen affective values were far from each other, and lower when they were close. Our result is in line with previous studies showing that MCC, and adjacent ACC, are found to be involved in the processing of affective primary value signals such as pain, and overall emotional experience (Peyron et al., 2000 [138]). Moreover, the anterior MCC is involved in regulating affective responses in general (Etkin et al., 2011 [164]; Shackman et al., 2011 [165]).

Taken together, our and others' data challenge the historical view of medial PFC as the more ventral the more affective, the more dorsal the more "cognitive" (Bush et al., 2000 [163]).

9.3 General Conclusion

In this PhD work, we provided experimental evidence that action outcomes convey two major types of value signals: belief values and affective values, which subsequently drive decision-making. Using behavioral paradigms, computational modeling and fMRI, we showed that both beliefs and affective values influenced subjects' choices, in a distinct manner.

In our fMRI paradigm, healthy human subjects had to decide between two shapes representing two underlying states, one of which was more frequently rewarded than the other one. The proposed rewards to obtain for each shape were displayed before each choice. Crucially, we manipulated the reward distributions underlying each shape to dissociate both signals. Logistic regressions analyses revealed that (1) Subjects extracted information from proposed rewards presented before choice, indicating that proposed rewards presented as cues influenced subjects' belief, rather than being processed purely as affective values and (2) Both beliefs and affective values influenced subjects' choices, but without expected values computation.

We built a computational model establishing the functional dissociation between beliefs and affective values. It integrates two parallel systems: (1) Reinforcement learning, dealing with affective values, and (2) Bayesian inference, dealing with beliefs. Importantly, the model describes decisions as a linear mixture of beliefs and affective values. It fitted behavioral data better and more parsimoniously than many other alternative models.

We then investigated how the belief and the affective value systems interact in the brain using fMRI. In particular, we tested whether beliefs and affective values have distinct neural bases. BOLD signal was regressed against linear and quadratic effects of both value signals, to dissociate between choice-dependent (linear) and choice-independent (quadratic) representations. We found choice-dependent representations in ventromedial prefrontal cortex and midcingulate cortex, for both beliefs and affective values. Presumably, they reflect expectations associated with chosen and unchosen shapes. Critically, we found a double-dissociation regarding choice-independent representations, with ventromedial prefrontal cortex encoding choice-independent preferences in terms of beliefs, whereas midcingulate cortex encoded choice-independent preferences in terms of affective values. Lastly, lateral prefrontal cortex showed a negative quadratic effect, meaning

that it was more activated when both decision values (i.e. combination of belief and affective value) were close to each other. Such a pattern suggests that lateral prefrontal cortex performs action selection.

A key feature of this PhD work lies in the complementary contribution of experimental and modeling approaches to the study of human decision-making. Taken together, these results suggest the following neural architecture underlying value-based decision-making in prefrontal cortex. Before decision, vmPFC and MCC separately encode beliefs and affective values respectively, as supported by the double-dissociation between vmPFC and MCC regarding choice-independent signals. Lateral PFC combines both signals to make a decision. In return, lateral PFC feeds back choice information to the medial regions, presumably for updating these value signals according to action outcomes, hence choice-dependent representations in both vmPFC and MCC. These results precise how the various prefrontal cortex subparts interact during human decision-making.

Appendix A

Informal debriefing for the first series of behavioral experiments

Remarques ?

Comment avez-vous trouvé l'expérience : long/court, facile/difficile ? perturbante ? Si oui, pourquoi ? Qu'est-ce qui vous a perturbé ?

Avez-vous utilisé une stratégie particulière pour répondre ?

Avez-vous repéré des régularités, une organisation particulière dans l'expérience ?

Avez-vous utilisé une règle en particulier après chaque type de récompense : 1, 2, 5, 8 ou 9 Euros ?

Lorsque vous pensiez avoir identifié la meilleure option, a quel type de récompense vous attendiez-vous ?

Y avait-il une valeur de récompense a partir de laquelle vous considériez que la tache que vous étiez en train de faire n'était plus la bonne ?

Quelles valeurs de récompense considériez-vous (virtuellement) équivalentes ?

Avez-vous repéré tous les combien d'essais, après combien d'essais la meilleure des deux options changeait ?

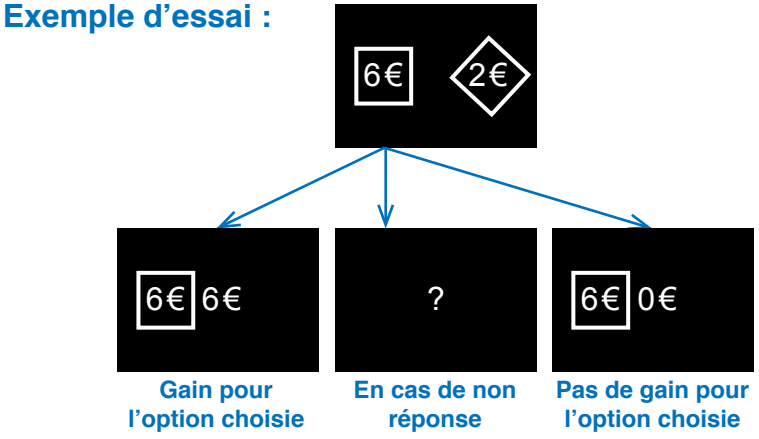
Autres remarques ?

Appendix B

Instructions for fMRI experiment

Consignes

Un losange et un carré s'affichent à l'écran. Au centre de chaque forme est indiquée la valeur que vous êtes susceptibles de gagner en choisissant cette forme. Vous devrez choisir l'une de ces deux options: Pour choisir l'option de gauche, appuyez sur la touche j. Pour choisir l'option de droite, appuyez sur la touche k. Une fois votre choix effectué, deux cas possibles : Soit vous gagnez la valeur indiquée dans la forme que vous avez choisie ; alors il apparaîtra au centre de l'écran le montant que vous remportez pour cet essai (1 à 11 Euros) Soit vous ne gagnez rien ; alors il apparaîtra au centre de l'écran 0 Euro Si vous ne répondez pas dans le temps imparti, l'essai est considéré comme perdu. N'appuyez sur une touche qu'une seule fois par essai : seule la première réponse est prise en compte.



L'une des deux formes conduit plus fréquemment à obtenir une récompense, mais celle-ci changera de temps en temps au cours de l'expérience. A vous faire le choix, à chaque essai, qui vous rapportera le plus d'argent.

Vous allez réaliser un entraînement de 5 minutes sur le jeu avant d'entrer dans l'IRM. Avez-vous des questions ?

Consignes supplémentaires

PAUSES L'expérience s'interrompra plusieurs fois au bout d'un certain temps pour vous laisser le temps de vous reposer. L'expérience se décompose en quatre blocs de 14 minutes chacun. Il n'y a aucun lien entre la survenue d'une pause et le déroulement de l'expérience. Après une pause, l'expérience reprend son cours, donc n'oubliez pas ce que vous faisiez avant la pause.

Vous devrez garder la tête fixe et immobile pendant chaque bloc. Nous mesurerons la position de votre tête dans le scanner au début de chaque bloc.

Merci de ne pas parler de l'expérience à des personnes qui la passeront après vous : nous nous intéressons à tester des personnes sans a priori sur l'expérience.

BONUS Votre bonus sera calculé sur la base de ce que vous gagnerez pendant l'expérience. Nous sélectionnerons au hasard 2% des essais de chaque session pour calculer votre bonus.

Avez vous envie de passer aux toilettes avant de commencer ?

Avez vous d'autres questions au sujet de l'expérience ?

Appendix C

Informal debriefing following the last fMRI session

Remarques ?

Comment avez-vous trouvé l'expérience : long/court, facile/difficile ? perturbante ? Si oui, pourquoi / qu'est-ce qui vous a perturbé ?

Avez-vous utilisé une stratégie particulière pour répondre ?

Avez-vous remarqué quelque chose de différent entre les 3 sessions ?

Avez-vous repéré des régularités, une organisation particulière dans l'expérience ?

Avez-vous utilisé une règle en particulier après chaque type de récompense : 1, . . . , 11 Euros ?

Avez-vous basé vos décisions plutôt sur les formes ou sur les valeurs affichées dans les formes ?

Quelles valeurs de récompense considériez-vous (virtuellement) équivalentes ?

Avez-vous repéré tous les combien d'essais, après combien d'essais la meilleure des deux options changeait ?

Avez-vous basé vos décisions sur les formes ou sur les valeurs ?

- Pour condition 1 :

- Pour condition 2 :

- Pour condition 3 :

Appendix D

Generative model of the fMRI task

The task aims at distinguishing how information and value are integrations, and how information-carrying value influences the subjects' behavior. The task corresponds to the Bayesian model discussed below. The text describes both Bayesian inference and a potential Reinforcement Learning model. This formal description of the model was written by Dr Jan Drugowitsch, who collaborated with us on the modeling part of the project.

D.1 Task Description

The task consists of a sequence of hidden states, $z_1, z_2, \ldots, z_t \in \{0, 1\}$, and it is the subject's task to maximize his/her overall reward. The generative model of the task is presented in Figure D.1. In each trial t the subject can choose between two options, 0 and 1, and it is the subject's task to choose the options that corresponds to the current hidden state, z_t. The subject receives information about the hidden state through two cues. On one hand, two rewards, $r_t^{(0)}$ and $r_t^{(1)}$ are presented to the subject before each choice. On the other hand, after the subject's choice, he/she receives feedback $x_t \in \{0, 1\}$ about the choice, which causes the reward associated with the choice to be given ($x_t = 1$) or not given ($x_t = 0$).

The structure that determines the above variables is as follows. The hidden state changes with volatility probability ν, such that

$$p(z_{t+1}|z_t) = \begin{cases} 1 - \nu & \text{if } z_{t+1} = z_t \text{(no switch)} \\ \nu & \text{otherwise (switch).} \end{cases} \tag{D.1}$$

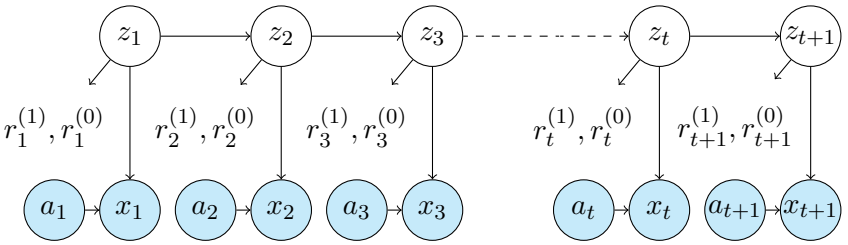


FIGURE D.1: Generative model of the task: z_t: underlying hidden state; $r_t^{(1)}, r_t^{(0)}$: proposed rewards before choice; a_t: action performed; x_t: feedback observed.

The rewards are drawn from some distribution $p(r_t^{(0)}|z_t)$ and $p(r_t^{(1)}|z_t)$, where $r_t^{(0)}$ is the reward for choosing the option associated with $z_t = 0$, and $r_t^{(1)}$ denotes the option associated with $z_t = 1$. A simple approach to specify this distribution is to choose

$$p(r_t^{(0)}|z_t) \propto \begin{cases} \exp(\gamma(r_t^{(0)} - \bar{r})) & \text{if } z_t = 0, \\ \exp(-\gamma(r_t^{(0)} - \bar{r})) & \text{otherwise,} \end{cases} \tag{D.2}$$

$$p(r_t^{(1)}|z_t) \propto \begin{cases} \exp(\gamma(r_t^{(1)} - \bar{r})) & \text{if } z_t = 1, \\ \exp(-\gamma(r_t^{(1)} - \bar{r})) & \text{otherwise,} \end{cases} \tag{D.3}$$

where $\bar{r}$ denotes the mean reward. In this case, $\gamma = 0$ indicates that the reward is uninformative about the hidden state, $\gamma > 0$ implies that choosing the high reward is better, and $\gamma < 0$ implies that choosing a low reward is better.

Once an action has been chosen, feedback is provided probabilistically according to

$$p(x_t = 1|z_t, a_t) = \begin{cases} q & \text{if } z_t = a_t \text{ (correct action)}, \\ 1 - q & \text{otherwise (wrong action)}, \end{cases} \tag{D.4}$$

and $p(x_t = 0|z_t, a_t) = 1 - p(x_t = 1|z_t, a_t)$. Thus, the task has the three parameters, ν, γ, and q.

D.2 Bayesian inference, known parameters

D.2.1 Inference

Assuming that the parameters ν, γ, and q are known, Bayesian inference of the hidden state is performed as follows. Assume we have inferred the belief $p(z_t = 1|r_{1:t-1}^{(0)}, r_{1:t-1}^{(1)}, a_{1:t-1}, x_{1:t-1})$, which equals $1 - p(z_t = 0|r_{1:t-1}^{(0)}, r_{1:t-1}^{(1)}, a_{1:t-1}, x_{1:t-1})$. We observe the proposed rewards $r_t^{(0)}$ and $r_t^{(1)}$ and want to update our belief accordingly. This update is performed by

Bayes rule, resulting in

$$\begin{aligned} p(z_t = 1|r^{(0)}_{1:t}, r^{(1)}_{1:t}, a_{1:t-1}, x_{1:t-1}) &\propto p(r^{(0)}_t|z_t = 1)p(r^{(1)}_t|z_t = 1) \\ &\quad \times p(z_t = 1|r^{(0)}_{1:t-1}, r^{(1)}_{1:t-1}, a_{1:t-1}, x_{1:t-1}) \\ &\propto e^{\gamma(r^{(1)}_t - r^{(0)}_t)} p(z_t = 1|r^{(0)}_{1:t-1}, r^{(1)}_{1:t-1}, a_{1:t-1}, x_{1:t-1}). \end{aligned} \tag{D.5}$$

The expression for $p(z_t = 0|\dots)$ is analogous, with the likelihood replaced by $e^{\gamma(r^{(0)}_t - r^{(1)}_t)}$. With both expressions at hand, these probabilities can be normalized to achieve $p(z_t = 1|\dots) + p(z_t = 0|\dots) = 1$. Note that the mean reward now becomes irrelevant, as the only thing that matters in the likelihood is the difference between the two observed rewards.

After either option has been chosen, feedback x_t is observed. This allows us to again update the belief by Bayes rule, resulting in

$$\begin{aligned} p(z_t = 1|r^{(0)}_{1:t}, r^{(1)}_{1:t}, a_{1:t}, x_{1:t}) &\propto p(x_t|z_t = 1, a_t)p(z_t = 1|r^{(0)}_{1:t}, r^{(1)}_{1:t}, a_{1:t-1}, x_{1:t-1}) \\ &\propto q^{x_t a_t + (1-x_t)(1-a_t)}(1-q)^{x_t(1-a_t)+(1-x_t)a_t} \\ &\quad \times p(z_t = 1|r^{(0)}_{1:t}, r^{(1)}_{1:t}, a_{1:t-1}, x_{1:t-1}), \end{aligned} \tag{D.6}$$

where the likelihood of $z_t = 1$ is q if either the feedback is positive ($x_t = 1$) if $a_t = 1$ has been choosen, or negative ($x_t = 0$) if $a_t = 0$ has been chosen, and $1 - q$ otherwise. The expression for $p(z_t = 0|\dots)$ is similar, with the likelihood replaced by $q^{x_t(1-a_t)+(1-x_t)a_t}(1-q)^{x_t a_t + (1-x_t)(1-a_t)}$, such that the belief is again easily normalised.

The belief computed so far is held at the end of trial t, after proposed rewards have been observed, an action has been performed, and feedback has been given. In order to compute the belief at the beginning of the next trial, $t + 1$, we need to use the belief transition mass function $p(z_{t+1}|z_t)$ to get

$$\begin{aligned} p(z_{t+1} = 1|r^{(0)}_{1:t}, r^{(1)}_{1:t}, a_{1:t}, x_{1:t}) &= \sum_{z_t \in \{0,1\}} p(z_{t+1} = 1|z_t)p(z_t|r^{(0)}_{1:t}, r^{(1)}_{1:t}, a_{1:t}, x_{1:t}) \\ &= (1-\nu)p(z_t = 1|r^{(0)}_{1:t}, r^{(1)}_{1:t}, a_{1:t}, x_{1:t}) \\ &\quad + \nu p(z_t = 0|r^{(0)}_{1:t}, r^{(1)}_{1:t}, a_{1:t}, x_{1:t}). \end{aligned} \tag{D.7}$$

An analogous expression holds for $p(z_{z+1} = 0|\dots)$.

D.2.2 Action selection

Optimal action selection in the Bayesian model depends on the expected reward for either option. For choosing option 1, for example, this option in according to the current belief

expected to be correct (that is, rewarded) with probability

$$\begin{aligned} p(x_t = 1|a_t, r^{(0)}_{1:t}, r^{(1)}_{1:t}, a_{1:t-1}, x_{1:t-1}) &= \sum_{z_t \in \{0,1\}} p(x_t = 1|a_t = 1, z_t) p(z_t|r^{(0)}_{1:t}, r^{(1)}_{1:t}, a_{1:t-1}, x_{1:t-1}) \\ &= qp(z_t = 1|r^{(0)}_{1:t}, r^{(1)}_{1:t}, a_{1:t-1}, x_{1:t-1}) \\ &\quad + (1-q)p(z_t = 0|r^{(0)}_{1:t}, r^{(1)}_{1:t}, a_{1:t-1}, x_{1:t-1}). \end{aligned} \tag{D.8}$$

A similar expression holds for choosing option 0. Thus, the expected rewards for choosing either option are

$$\langle r \rangle_{a_t=0} = r^{(0)}_t \left(q(1 - p(z_t = 1|\dots)) + (1-q)p(z_t = 1|\dots) \right), \tag{D.9}$$

$$\langle r \rangle_{a_t=1} = r^{(1)}_t \left(qp(z_t = 1|\dots) + (1-q)(1 - p(z_t = 1|\dots)) \right). \tag{D.10}$$

The subject chooses optimally if

$$a_t = \begin{cases} 1 & \text{if } \langle r \rangle_{a_t=1} > \langle r \rangle_{a_t=0}, \\ 0 & \text{otherwise.} \end{cases} \tag{D.11}$$

An alternative action selection strategy is to perform stochastic action selection according to

$$p(a_t = 1) = \frac{\epsilon}{2} + (1-\epsilon) \frac{e^{\beta \langle r \rangle_{a_t=1}}}{e^{\beta \langle r \rangle_{a_t=0}} + e^{\beta \langle r \rangle_{a_t=1}}}, \tag{D.12}$$

with parameters β and ϵ.

D.2.3 Contributions to action selection

Assuming stochastic action selection with $\epsilon \approx 0$, we can write the action probability in terms of the log-odds $\ell_t = \beta(\langle r \rangle_{a_t=1} - \langle r \rangle_{a_t=0})$ as

$$p(a_t = 1) = \frac{1}{1 + e^{-\ell_t}}, \tag{D.13}$$

which is the logistic sigmoid. This allows us to use logistic regression to test the predictions of the Bayesian model on how observables influence action selection.

At first, let us substitute for the $\langle r \rangle$'s to find

$$\ell_t = \beta(1-q) r^{(1)}_t - \beta q r^{(0)}_t + \beta(r^{(0)}_t + r^{(1)}_t)(2q-1) p(z_t = 1|\dots), \tag{D.14}$$

with partial derivatives

$$\frac{\partial \ell_t}{\partial r_t^{(0)}} = \beta(1-q) + \beta(2q-1)p(z_t = 1|\dots) > 0, \tag{D.15}$$

$$\frac{\partial \ell_t}{\partial r_t^{(1)}} = -\beta q + \beta(2q-1)p(z_t = 1|\dots) < 0, \tag{D.16}$$

$$\frac{\partial \ell_t}{\partial p(z_t = 1|\dots)} = \beta(r_t^{(0)} + r_t^{(1)})(2q-1) > 0, \tag{D.17}$$

where the inequalities are due to $0 \leq p(z_t|\dots) \leq 1$, $q > \frac{1}{2}$, and $r_t^{(0)} + r_t^{(1)} > 0$. This shows that the log-odds of choosing action 1 increases strictly with $r_t^{(1)}$, decreases strictly with $r_t^{(0)}$, and increases strictly with $p(z_t = 1|\dots)$. Note, however, that these relations do not take into account how the observed reward influences $p(z_t = 1|\dots)$.

To determine the influence of rewards on $p(z_t = 1|\dots)$, let us first observe that

$$p(z_t = 1|\dots) = \frac{e^{\gamma \Delta_{rt}} \tilde{p}_t}{e^{\gamma \Delta_{rt}} \tilde{p}_t + e^{-\gamma \Delta_{rt}}(1-\tilde{p}_t)} = \frac{1}{1 + e^{-(2\gamma \Delta_{rt} + \log \frac{\tilde{p}_t}{1-\tilde{p}_t})}}, \tag{D.18}$$

where $\Delta_{rt} = r_t^{(1)} - r_t^{(0)}$ and $\tilde{p}_t = p(z+t = 1|r_{1:t-1}^{(0)}, r_{1:t-1}^{(1)}, a_{1:t-1}, x_{1:t-1})$. This shows that $p(z_t = 1|\dots)$ is a logistic sigmoid in $2\gamma\Delta_{rt} + \log \frac{\tilde{p}_t}{1-\tilde{p}_t}$, and therefore strictly increasing in this quantity. As a result, $p(z_t = 1|\dots)$ increases if $\gamma(r_t^{(1)} - r_t^{(0)}) > 0$ and decreases otherwise. Thus, in the case of $\gamma > 0$ it is increasing in $r_t^{(1)}$ and decreasing in $r_t^{(0)}$, consistent with the log-odds ℓ_t. On the other hand, if $\gamma < 0$, then it is decreasing in $r_t^{(0)}$ and increasing in $r_t^{(1)}$, opposing the direct dependency of ℓ_t on the reward. This shows that the currently observed rewards influence action selection both when updating the belief and when computing the expected reward. The direction in which they influence the belief update depends on the sign of γ. With respect to expected reward, they influence the log-odds always increasingly in $r_t^{(1)}$ and decreasingly in $r_t^{(0)}$.

This analysis also reveals that ℓ_t is strictly increasing in $p(z_t = 1|\dots)$, which in turn is by Eq. (D.18) increasing in the belief $\tilde{p}_t$ before the current rewards have been observed. This monotonicity is preserved by the transition step Eq. (D.7), a fact we use as basis to investigate how reward and feedback in previous trials influences action selection in the current trial.

Considering first feedback, Eq. (D.6) can be re-written as the sigmoid

$$\left(1 + e^{-\left(\mathcal{I}(a_{t-1}=x_{t-1}) \log \frac{q}{1-q} + \mathcal{I}(a_{t-1}\neq x_{t-1}) \log \frac{1-q}{q} + \log \frac{p(z_{t-1}=1|\dots)}{1-p(z_{t-1}|\dots)}\right)}\right)^{-1}, \tag{D.19}$$

where $\mathcal{I}(a)$ is the identifier function that returns $\mathcal{I}(a) = 1$ if a is true and $\mathcal{I}(a) = 0$ otherwise. Observing that due to $q > \frac{1}{2}$ we have $\log \frac{q}{1-q} > 0$ and $\log \frac{1-q}{q} < 0$, we can see

that if action $a_{t-1} = 1$ was chosen, positive past feedback, $x_{t-1} = 1$ increases ℓ_t, whereas negative past feedback, $x_{t-1} = 0$ decreases ℓ_t. Analogously, if action $a_{t-1} = 0$ was chosen, negative past feedback, $x_{t-1} = 0$ increases ℓ_t, whereas positive past feedback, $x_{t-1} = 1$ decreases ℓ_t, as one would intuitively expect. Due to the monotonicity of sequential belief updates, this also applies to all feedbacks given before $t-1$. Thus, qualitatively, the log-odds, ℓ_t is increasing in $\eta_{x,t-n}(-1)^{\mathcal{I}(a_{t-n} \neq x_{t-n})}$ for all $n > 0$, where $\eta_{x,t-n}$ is some positive scalar.

The influence of past reward on ℓ_t is derived similarly. First, we re-write Eq. (D.5) for trial $t-1$ as the logistic sigmoid

$$\left(1 + e^{-\left(2\gamma\Delta_{rt-1} + \log \frac{\tilde{p}_{t-1}}{1-\tilde{p}_{t-1}}\right)}\right)^{-1}, \tag{D.20}$$

which is increasing in $\gamma\Delta_{rt-1} = \gamma(r^{(1)}_{t-1} - r^{(0)}_{t-1})$. Therefore, if $\gamma > 0$, ℓ_t is increasing in $r^{(1)}_{t-1}$ and decreasing in $r^{(0)}_{t-1}$. Conversely, if $\gamma < 1$, ℓ_t is decreasing in $r^{(1)}_{t-1}$ and increasing in $r^{(0)}_{t-1}$. Due to the monotonic belief update, this also applies to reward past $t-1$. Thus, qualitatively, the log-odds, ℓ_t, is increasing in $\eta_{r,t-n}(-1)^{\mathcal{I}(\gamma<0)}(r^{(1)}_{t-n} - r^{(0)}_{t-n})$ for all $n > 0$, where $\eta_{r,t-n}$ is again some positive scalar.

In summary, we can write the log-posterior approximately as the sequence

$$\ell_t = -\eta_0 r^{(0)}_t + \eta_1 r^{(1)}_t + \sum_{n=1}^{t-1} \eta_{x,t-n}(-1)^{\mathcal{I}(a_{t-n} \neq x_{t-n})} + (-1)^{\mathcal{I}(\gamma<0)} \sum_{n=1}^{t-1} \eta_{r,t-n}(r^{(1)}_{t-n} - r^{(0)}_{t-n}), \tag{D.21}$$

where all $\eta_{x,t-n}$'s and $\eta_{r,t-n}$'s are positive, and η_0 and η_1 are positive if $\gamma > 0$ and might be either positive or negative if $\gamma < 0$.

D.3 Reinforcement learning

D.3.1 Inference

Assume that we maintain the expected returns in $Q^{(0)}_t$ for choice 0, and in $Q^{(1)}_t$ for choice 1. Reinforcement learning does not provide any mechanism per se to include promised rewards, as provided in our task. Thus, they will only considered during action selection.

To update the expected returns, we combine both reward and feedback, and update only the expected return of the chosen action. This return is updated with the given reward if the feedback was positive, or reward 0 if the feedback was negative. This results in

the update equation

$$Q_{t+1}^{(j)} = \begin{cases} Q_t^{(j)} + \alpha(x_t r_t^{(j)} - Q_t^{(j)}) & \text{if } a_t = j, \\ Q_t^{(j)} & \text{otherwise,} \end{cases} \tag{D.22}$$

for $j \in \{0, 1\}$, with learning rate α.

D.3.2 Action selection

We could perform action selection by ϵ-softmax on the current expected returns. However, this option would not take into account information about the rewards that are currently displayed. In order to consider these, we can bias the current expected returns by $\tilde{Q}_t^{(j)} = Q_t^{(j)} + w(r_t^{(j)} - Q_t^{(j)})$, and then perform the action based on these biased returns. This results in the stochastic choice

$$p(a_t = 1) = \frac{\epsilon}{2} + (1 - \epsilon)\frac{e^{\beta((1-w)Q_t^{(1)} + wr_t^{(1)})}}{e^{\beta((1-w)Q_t^{(0)} + wr_t^{(0)})} + e^{\beta((1-w)Q_t^{(1)} + wr_t^{(1)})}}, \tag{D.23}$$

with parameters ϵ, β, and w. The bias corresponds to a "what if" forward step that is only performed for action selection and not taken into account for inference.

D.3.3 Contributions to action selection

To decompose the action probability, we assume $\epsilon \approx 0$ such that the log-odds, $\ell_t = \log(p(a_t = 1)/p(a_t = 0))$ can be written as

$$\ell_t = \beta w r_t^{(1)} - \beta w r_t^{(0)} + \beta(1 - w)Q_t^{(1)} - \beta(1 - w)Q_t^{(0)}. \tag{D.24}$$

Using Eq. (D.22) to substitute for $Q_t^{(0)}$ and $Q_t^{(1)}$ we find that these log-odds can be expressed by

$$\begin{aligned} \ell_t \;=\; & \beta w r_t^{(1)} - \beta w r_t^{(0)} \\ & + (-1)^{1-a_{t-1}} \beta(1 - w)\alpha x_{t-1} r_{t-1}^{(a_{t-1})} \\ & + \beta(1 - w)(1 - \alpha)^{a_{t-1}} Q_{t-1}^{(1)} - \beta(1 - w)(1 - \alpha)^{1-a_{t-1}} Q_{t-1}^{(0)}. \end{aligned} \tag{D.25}$$

This shows that, according to reinforcement learning, only the reward $r^{(a_{t-1})}$ corresponding to the previously chosen option a_{t-1} influences the current choice a_t. Furthermore, by the pre-factor $(-1)^{1-a_{t-1}}$ it influences the log-odds positively if $a_{t-1} = 1$, and negatively otherwise. The discounting of the previous Q-values also depends on the choice in the previous trial.

We go a further step backwards in time by taking the above expression and re-substituting Eq. (D.22) for $Q_{t-1}^{(0)}$ and $Q_{t-2}^{(1)}$. This yields the more complex expression

$$\begin{aligned}\ell_t &= \beta w r_t^{(1)} - \beta w r_t^{(0)} \\ &\quad +(-1)^{1-a_{t-1}}\beta(1-w)\alpha x_{t-1} r_{t-1}^{(a_{t-1})} \\ &\quad +(-1)^{1-a_{t-2}}\beta(1-w)(1-\alpha)^{a_{t-2}a_{t-1}+(1-a_{t-2})(1-a_{t-1})}\alpha x_{t-2} r_{t-2}^{(a_{t-2})} \\ &\quad +\beta(1-w)(1-\alpha)^{a_{t-1}+a_{t+2}}Q_{t-2}^{(1)} - \beta(1-w)(1-\alpha)^{(1-a_{t-1})+(1-a_{t-2})}Q_{t-2}^{(0)},\end{aligned} \tag{D.26}$$

which shows a qualitatively similar dependency on past reward. If we continue re-substituting Eq. (D.22) we reach the final expression, which is given by

$$\begin{aligned}\ell_t &= \beta w r_t^{(1)} - \beta w r_t^{(0)} \\ &\quad +\beta(1-w)\alpha\sum_{n=1}^{t-1}(-1)^{1-a_{t-n}}(1-\alpha)^{a_{t-n}\sum_{j=1}^{n-1}a_{t-j}+(1-a_{t-n})\sum_{j=1}^{n-1}(1-a_{t-j})}x_{t-n}r_{t-n}^{(a_{t-n})} \\ &\quad +\beta(1-w)(1-\alpha)^{\sum_{j=1}^{t-1}a_{t-j}}Q_1^{(1)} + \beta(1-w)(1-\alpha)^{\sum_{j=1}^{t-1}(1-a_{t-j})}Q_1^{(0)},\end{aligned} \tag{D.27}$$

where $Q_1^{(0)}$ and $Q_1^{(1)}$ are the initial Q-values. Note that the complex pre-factor to $x_{t-n}r_{t-n}^{(a_{t-n})}$ is required to take into account the whole sequence of α-weighted updates in Eq. (D.22).

Overall, if we ignore the influence of the initial Q-values, the log-odds is decomposed into

$$\ell_t = \beta w r_t^{(1)} - \beta w r_t^{(0)} + \sum_{n=1}^{t-1}\eta_{t-n}(-1)^{1-a_{t-n}}x_{t-n}r_{t-n}^{(a_{t-n})}, \tag{D.28}$$

with non-negative contribution weights

$$\eta_{t-n} = \beta(1-w)\alpha(1-\alpha)^{a_{t-n}\sum_{j=1}^{n-1}a_{t-j}+(1-a_{t-n})\sum_{j=1}^{n-1}(1-a_{t-j})}, \tag{D.29}$$

that depend on the sequence of chosen actions. This shows again that, according to the reinforcement learning model, only rewards associated with the chosen action should have an influence on future action selection.

We used logistic regressions to test the influence of the various protocol variables on choice.

Formally, the Bayesian model predicted that choice should only depend on the following variables: $r_t^{(1)}$ and $r_t^{(0)}$, the proposed rewards before choice at time t, all pairs of proposed rewards before choice at all previous trials, $r_{t-n}^{(1)}$ and $r_{t-n}^{(0)}$, previous binary feedback obtained (win or loose), x_{t-n}, and the initial beliefs (mathematical demonstration is described in Appendix D). A pure model-free logistic regression (Figure D.2) showed a

contribution of reward received at previous trial to choice (last bar in condition anti-correlated), after accounting for all the above variables (up to *t-1*). This significant effect of previous reward obtained violated the Bayesian model predictions, in which no explicit influence of previous reward was included (Figure D.2).

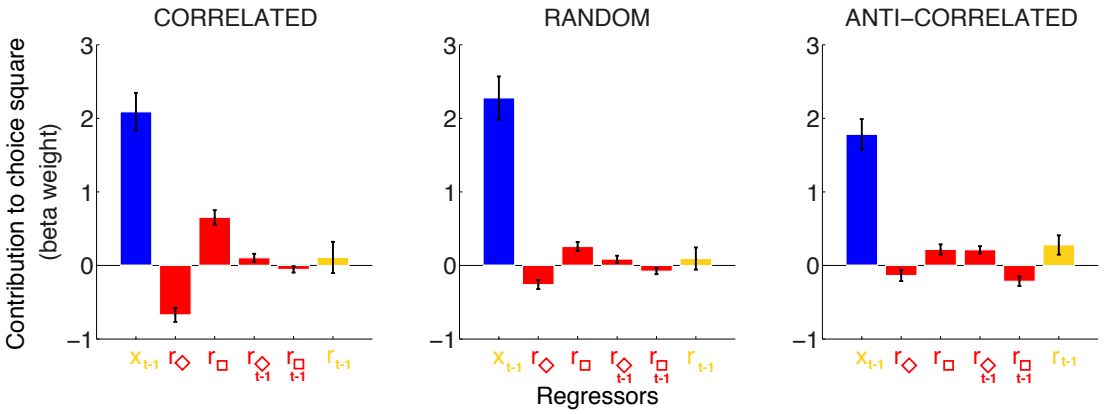


FIGURE D.2: Logistic regression violating the predictions of a pure Bayesian model.

Formally, deriving the Standard RL model (details in Appendix D), choice should depend only on $r_t^{(1)}$ and $r_t^{(0)}$, the rewards presented before choice, on all the previous received rewards and on initial Q values. In the condition anti-correlated, we found a significant contribution of unchosen previous reward to current choice (last bar), even when accounting for all the above variables (up to *t-1*). This formally violated pure RL predictions (Figure D.3).

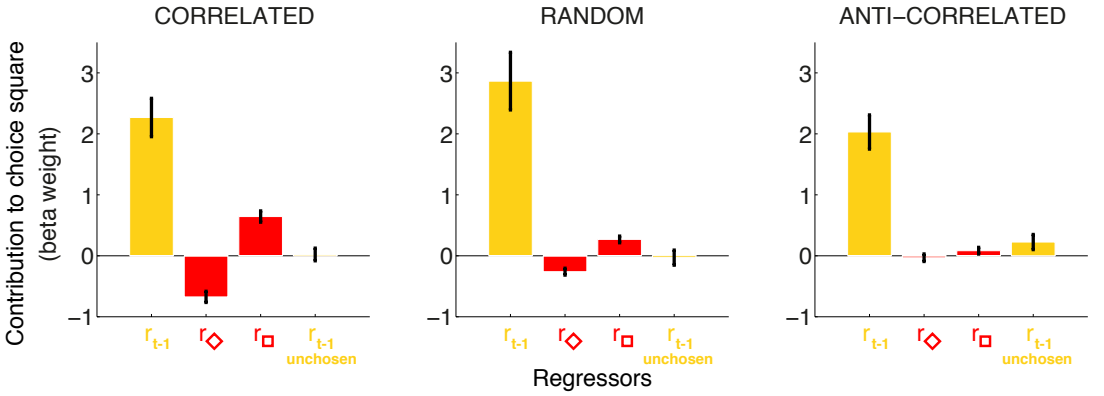


FIGURE D.3: Logistic regression violating the predictions of a pure RL model.

Bibliography

[1] Earl K Miller and Jonathan D Cohen. An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, 24(1):167–202, 2001.

[2] Hanna Damasio, Thomas Grabowski, Randall Frank, Albert M Galaburda, and Antonio R Damasio. The Return of Phineas Gage: Clues About the Brain from the Skull of a Famous Patient. *Science*, 264:1102–1105, May 1994.

[3] Tim Shallice and Paul Burgess. Deficits in strategy application following frontal lobe damage in man. *Brain*, pages 1–15, November 1991.

[4] Joshua Rubinstein, Jeffrey E Evans, and David E Meyer. Task switching in patients with prefrontal cortex damage. In *annual meeting of the Cognitive Neuroscience Society, San Francisco, CA*, 1994.

[5] Sara M Szczepanski and Robert T Knight. Insights into Human Behavior from Lesions to the Prefrontal Cortex. *Neuron*, 83(5):1002–1018, September 2014.

[6] Carole Azuar, Pablo Reyes, Andrea Slachevsky, Emmanuelle Volle, Serge Kinkingnehun, Frédérique Kouneiher, Eduardo Bravo, Bruno Dubois, Etienne Koechlin, and Richard Levy. Testing the model of caudo-rostral organization of cognitive control in the human with frontal lesions. *NeuroImage*, 84(C):1053–1060, January 2014.

[7] Louis N Irwin. Comparative Neuroscience and Neurobiology. In Springer, editor, *Encyclopedia of Neuroscience*, pages 1–146. November 1988.

[8] Anne G E Collins and Michael J Frank. How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience*, 35(7):1024–1035, April 2012.

[9] Michael Petrides and D N Pandya. Comparative cytoarchitectonic analysis of the human and the macaque ventrolateral prefrontal cortex and corticocortical connection patterns in the monkey. *European Journal of Neuroscience*, 16:291–310, May 2002.

[10] Paula L Croxson, Heidi Johansen-Berg, Timothy E J Behrens, Matthew D Robson, Mark A Pinsk, Charles G Gross, Wolfgang Richter, Marlene C Richter, Sabine Kastner, and Matthew F S Rushworth. Quantitative investigation of connections of the prefrontal cortex in the human and macaque using probabilistic diffusion tractography. *Journal of Neuroscience*, 25(39):8854–8866, September 2005.

[11] Steven P Wise. Forward frontal fields: phylogeny and fundamental function. *Trends in Neurosciences*, 31(12):599–608, December 2008.

[12] Todd M Preuss. Do rats have prefrontal cortex? The Rose-Woolsey-Akert program reconsidered. *Journal of Cognitive Neuroscience*, 7(1):1–24, 1995.

[13] Thomas A Stalnaker, Nisha K Cooch, and Geoffrey Schoenbaum. What the orbitofrontal cortex does not do. *Nature Neuroscience*, 18(5):620–627, April 2015.

[14] Emmanuel Procyk, Charles R E Wilson, Frederic M Stoll, Maïlys C M Faraut, Michael Petrides, and Céline Amiez. Midcingulate Motor Map and Feedback Detection: Converging Data from Humans and Monkeys. *Cerebral cortex (New York, N.Y. : 1991)*, pages 1–10, September 2014.

[15] Michael Petrides, Francesco Tomaiuolo, Edward H Yeterian, and Deepak N Pandya. The prefrontal cortex: comparative architectonic organization in the human and the macaque monkey brains. *Cortex*, 48(1):46–57, 2012.

[16] Céline Amiez, Rémi Neveu, Delphine Warrot, Michael Petrides, Kenneth Knoblauch, and Emmanuel Procyk. The location of feedback-related activity in the midcingulate cortex is predicted by local morphology. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 33(5):2217–2228, January 2013.

[17] Brent A Vogt. Architecture, neurocytology and comparative organization of monkey and human cingulate cortices. *Cingulate neurobiology and disease*, pages 65–93, 2009.

[18] Francesca Bonini, Boris Burle, Catherine Liegeois-Chauvel, Jean Regis, Patrick Chauvel, and Franck Vidal. Action Monitoring and Medial Frontal Cortex: Leading Role of Supplementary Motor Area. *Science*, 343(6173):888–891, February 2014.

[19] Michael S Gazzaniga. *The Cognitive Neurosciences*. MIT PRESS, September 2009.

[20] Sarah-Jayne Blakemore. The Developing Social Brain: Implications for Education. *Neuron*, 65(6):744–747, March 2010.

[21] Sarah-Jayne Blakemore. The social brain in adolescence. *Nature Reviews Neuroscience*, 9(4):267–277, April 2008.

[22] Iroise Dumontheil, Rachael Houlton, Kalina Christoff, and Sarah-Jayne Blakemore. Development of relational reasoning during adolescence. *Developmental Science*, 13(6):F15–F24, October 2010.

[23] Samuel R Chamberlain, Lara Menzies, Adam Hampshire, John Suckling, Naomi A Fineberg, Natalia del Campo, Mike Aitken, Kevin Craig, Adrian M Owen, Edward T Bullmore, Trevor W Robbins, and Barbara J Sahakian. Orbitofrontal Dysfunction in Patients with Obsessive-Compulsive Disorder and Their Unaffected Relatives . *Science*, 321(5887):421–422, July 2008.

[24] Lewis R Baxter, Jeffrey M Schwartz, Kenneth S Bergman, Martin P Szuba, Barry H Guze, John C Mazziotta, Adina Alazraki, Carl E Selin, Huan-Kwang Ferng, and Paul Munford. Caudate glucose metabolic rate changes with both drug and behavior therapy for obsessive-compulsive disorder. *Archives of General Psychiatry*, 49(9):681–689, 1992.

[25] Rita Z Goldstein and Nora D Volkow. Dysfunction of the prefrontal cortex in addiction: neuroimaging findings and clinical implications. *Nature Reviews Neuroscience*, 12(652):1–18, November 2011.

[26] Ruben D Baler and Nora D Volkow. Drug addiction: the neurobiology of disrupted self-control. *Trends in molecular medicine*, 12(12):559–566, 2006.

[27] Nigel J Blackwood, Robert J Howard, Richard P Bentall, and Robin M Murray. Cognitive Neuropsychiatric Models of Persecutory Delusions. *American Journal of Psychiatry*, 158(4):527–539, March 2001.

[28] Steffen Moritz and Todd S Woodward. Jumping to conclusions in delusional and non-delusional schizophrenic patients. *British Journal of Clinical Psychology*, 44 (2):193–207, January 2005.

[29] Renaud Jardri and Sophie Deneve. Circular inferences in schizophrenia. *Brain*, 136(11):3227–3241, October 2013.

[30] Paul C Fletcher and Chris D Frith. Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, 10(1):48–58, December 2008.

[31] Todd S Woodward, Steffen Moritz, Mahesh Menon, and Ruth Klinge. Belief inflexibility in schizophrenia. *Cognitive Neuropsychiatry*, 13(3):267–277, May 2008.

[32] Guillaume Barbalat, Marion Rouault, Narges Bazargani, Sukhwinder Shergill, and Sarah-Jayne Blakemore. The influence of prior expectations on facial expression discrimination in schizophrenia. *Psychological Medicine*, 42(11):2301–2311, March 2012.

[33] Valerian Chambon, Elisabeth Pacherie, Guillaume Barbalat, Pierre Jacquet, Nicolas Franck, and Chlöé Farrer. Mentalizing under influence: abnormal dependence on prior expectations in patients with schizophrenia. *Brain*, 134:3728–3741, 2011.

[34] Guillaume Barbalat, Valerian Chambon, Nicolas Franck, Etienne Koechlin, and Chlöé Farrer. Organization of Cognitive Control Within the Lateral Prefrontal Cortex in Schizophrenia. *Arch Gen Psychiatry*, 66(4):377–386, 2009.

[35] Guillaume Barbalat, Valérian Chambon, Philippe J D Domenech, Chrystèle Ody, Etienne Koechlin, Nicolas Franck, and Chloe Farrer. Impaired Hierarchical Control Within the Lateral Prefrontal Cortex in Schizophrenia. *BPS*, 70(1):73–80, July 2011.

[36] Etienne Koechlin, Chrystèle Ody, and Frédérique Kouneiher. The Architecture of Cognitive Control in the Human Prefrontal Cortex. *Science*, 302(5648):1181–1185, November 2003.

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

[38] Christian C Ruff, Giuseppe Ugazio, and Ernst Fehr. Changing Social Norm Compliance with Noninvasive Brain Stimulation. *Science*, 342(6157):482–484, October 2013.

[39] Katsuyuki Sakai and Richard E Passingham. Prefrontal interactions reflect future task operations. *Nature Neuroscience*, 6(1):75–81, January 2003.

[40] Katsuyuki Sakai and Richard E Passingham. Prefrontal set activity predicts rule-specific neural processing during subsequent cognitive performance. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 26(4):1211–1218, January 2006.

[41] Silvia A Bunge, Jonathan D Wallis, Amanda Parker, Marcel Brass, Eveline A Crone, Eiji Hoshi, and Katsuyuki Sakai. Neural Circuitry Underlying Rule Use in Humans and Nonhuman Primates. *Journal of Neuroscience*, 25(45):10347–10350, November 2005.

[42] Aldo Genovesio, Peter J Brasted, Andrew R Mitz, and Steven P Wise. Prefrontal Cortex Activity Related to Abstract Response Strategies. *Neuron*, 47(2):307–320, July 2005.

[43] Alan Baddeley. Working memory. *Current Biology*, 20(4):R136–R140, February 2010.

[44] Klaus Oberauer. Access to information in working memory: exploring the focus of attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(3):411, 2002.

[45] Klaus Oberauer and Reinhold Kliegl. A formal model of capacity limits in working memory. *Journal of Memory and Language*, 55(4):601–626, October 2006.

[46] Richard Lévy and Patricia S Goldman Rakic. Association of Storage and Processing Functions in the Dorsolateral Prefrontal Cortex of the Nonhuman Primate. *Journal of Neuroscience*, 19(12):5149–5158, June 1999.

[47] Joaqun M Fuster. The prefrontal cortex—an update: time is of the essence. *Neuron*, 30(2):319–333, 2001.

[48] Camillo Padoa-Schioppa and John A Assad. Neurons in the orbitofrontal cortex encode economic value. *Nature*, 441(7090):223–226, April 2006.

[49] Mael Lebreton, Soledad Jorge, Vincent Michel, Bertrand Thirion, and Mathias Pessiglione. An Automatic Valuation System in the Human Brain: Evidence from Functional Neuroimaging. *Neuron*, 64(3):431–439, November 2009.

[50] Charlotte Prévost, Mathias Pessiglione, Elise Météreau, Marie-Laure Cléry-Melin, and Jean-Claude Dreher. Separate valuation subsystems for delay and effort decision costs. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 30(42):14080–14090, October 2010.

[51] John A Clithero and Antonio Rangel. Informatic parcellation of the network involved in the computation of subjective value. *Social Cognitive and Affective Neuroscience*, page nst106, 2013.

[52] Vikram S Chib, Antonio Rangel, Shinsuke Shimojo, and John P O'Doherty. Evidence for a Common Representation of Decision Values for Dissimilar Goods in Human Ventromedial Prefrontal Cortex. *Journal of Neuroscience*, 29(39):12315–12320, September 2009.

[53] Hilke Plassmann, John P O'Doherty, and Antonio Rangel. Orbitofrontal Cortex Encodes Willingness to Pay in Everyday Economic Transactions. *Journal of Neuroscience*, 27(37):9984–9988, September 2007.

[54] John P O'Doherty. The problem with value. *Neuroscience and Biobehavioral Reviews*, 43:259–268, June 2014.

[55] Ryan K Jessup and John P O'Doherty. Distinguishing informational from value-related encoding of rewarding and punishing outcomes in the human brain. *European Journal of Neuroscience*, 39(11):2014–2026, May 2014.

[56] Alan N Hampton, Peter Bossaerts, and John P O'Doherty. The Role of the Ventromedial Prefrontal Cortex in Abstract State-Based Inference during Decision Making in Humans. *Journal of Neuroscience*, 26(32):8360–8367, August 2006.

[57] Joshua L Jones, Guillem R Esber, Michael A McDannald, Aaron J Gruber, Alex Hernandez, Aaron Mirenzi, and Geoffrey Schoenbaum. Orbitofrontal Cortex Supports Behavior and Learning Using Inferred But Not Cached Values. *Science*, 338 (6109):953–956, November 2012.

[58] Mathieu Roy, Daphna Shohamy, and Tor D Wager. Ventromedial prefrontal-subcortical systems and the generation of affective meaning. *Trends in Cognitive Sciences*, 16(3):147–156, March 2012.

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

[60] Mark E Walton, Timothy E J Behrens, MaryAnn P Noonan, and Matthew F S Rushworth. Giving credit where credit is due: orbitofrontal cortex and valuation in an uncertain world. *Annals of the New York Academy of Sciences*, 1239(1): 14–24, December 2011.

[61] Peter H Rudebeck and Elisabeth A Murray. The Orbitofrontal Oracle: Cortical Mechanisms for the Prediction and Evaluation of Specific Behavioral Outcomes. *Neuron*, 84(6):1143–1156, 2014.

[62] Robert C Wilson, Yuji K Takahashi, Geoffrey Schoenbaum, and Yael Niv. Orbitofrontal Cortex as a Cognitive Map of Task Space. *Neuron*, 81(2):267–279, January 2014.

[63] Brent A Vogt, Leslie Vogt, Nuri B Farber, and George Bush. Architecture and neurocytology of monkey cingulate gyrus. *The Journal of Comparative Neurology*, 485(3):218–239, 2005.

[64] Orrin Devinsky, Martha J Morrell, and Brent A Vogt. Contributions of anterior cingulate cortex to behaviour. *Brain*, 118:279–306, June 1995.

[65] William J Gehring, Brian Goss, Michael G H Coles, David E Meyer, and Emanuel Donchin. A neural system for error detection and compensation. *Psychological Science*, 4(6):385–390, November 1993.

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

[67] Michael J Frank, Brion S Woroch, and Tim Curran. Error-Related Negativity Predicts Reinforcement Learning and Conflict Biases. *Neuron*, 47(4):495–501, August 2005.

[68] Martin J Herrmann, Josefine Römmler, Ann-Christine Ehlis, Anke Heidrich, and Andreas J Fallgatter. Source localization (LORETA) of the error-related-negativity (ERN/Ne) and positivity (Pe). *Cognitive Brain Research*, 20(2):294–299, July 2004.

[69] Clémence Roger, Christian G Bénar, Franck Vidal, Thierry Hasbroucq, and Boris Burle. Rostral Cingulate Zone and correct response monitoring: ICA and source localization evidences for the unicity of correct- and error-negativities. *NeuroImage*, 51(1):391–403, May 2010.

[70] Laure Spieser, Wery van den Wildenberg, Thierry Hasbroucq, K Richard Ridderinkhof, and Boris Burle. Controlling your impulses: electrical stimulation of the human supplementary motor complex prevents impulsive errors. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 35(7): 3010–3015, February 2015.

[71] Matthew M Botvinick, Todd S Braver, Deanna M Barch, Cameron S Carter, and Jonathan D Cohen. Conflict monitoring and cognitive control. *Psychological Review*, 108(3):624, 2001.

[72] Matthew M Botvinick, Jonathan D Cohen, and Cameron S Carter. Conflict monitoring and anterior cingulate cortex: an update. *Trends in Cognitive Sciences*, 8 (12):539–546, December 2004.

[73] Amitai Shenhav, Mark A Straccia, Jonathan D Cohen, and Matthew M Botvinick. Anterior cingulate engagement in a foraging context reflects choice difficulty, not foraging value. *Nature Publishing Group*, 17(9):1249–1254, July 2014.

[74] Deanna M Barch, Todd S Braver, Erbil Akbudak, Tom Conturo, John Ollinger, and Avraham Snyder. Anterior Cingulate Cortex and Response Conflict: Effects of Response Modality and Processing Domain. *Cerebral cortex (New York, N.Y. : 1991)*, 11:837–848, August 2001.

[75] Boris Burle, Clémence Roger, Sonia Allain, Franck Vidal, and Thierry Hasbroucq. Error Negativity Does Not Reflect Conflict: A Reappraisal of Conflict Monitoring and Anterior Cingulate Cortex Activity. *Journal of Cognitive Neuroscience*, 20 (9):1637–1655, July 2008.

[76] Boris Burle, Sonia Allain, Franck Vidal, and Thierry Hasbroucq. Sequential Compatibility Effects and Cognitive Control: Does Conflict Really Matter? *Journal of Experimental Psychology: Human Perception and Performance*, 31(4):831–837, 2005.

[77] Frédérique Kouneiher, Sylvain Charron, and Etienne Koechlin. Motivation and cognitive control in the human prefrontal cortex. *Nature Publishing Group*, 12(7): 939–945, June 2009.

[78] Simon B Eickhoff, Angela R Laird, Peter T Fox, Danilo Bzdok, and Lukas Hensel. Functional Segregation of the Human Dorsomedial Prefrontal Cortex. *Cerebral cortex (New York, N.Y. : 1991)*, pages 1–18, October 2014.

[79] Pyungwon Kang, Jongbin Lee, Sunhae Sul, and Hackjin Kim. Dorsomedial prefrontal cortex activity predicts the accuracy in estimating others' preferences. *Frontiers in human neuroscience*, 7:1–11, November 2013.

[80] Adam Waytz, Jamil Zaki, and Jason P Mitchell. Response of Dorsomedial Prefrontal Cortex Predicts Altruistic Behavior. *Journal of Neuroscience*, 32(22):7646–7650, May 2012.

[81] Benjamin Y Hayden, John M Pearson, and Michael L Platt. Neuronal basis of sequential foraging decisions in a patchy environment. *Nature Publishing Group*, 14(7):933–939, June 2011.

[82] Nicholas Furl and Bruno B Averbeck. Parietal Cortex and Insula Relate to Evidence Seeking Relevant to Reward-Related Decisions. *Journal of Neuroscience*, 31(48):17572–17582, November 2011.

[83] Nils Kolling, Timothy E J Behrens, Rogier B Mars, and Matthew F S Rushworth. Neural Mechanisms of Foraging. *Science*, 336(6077):95–98, April 2012.

[84] Kenji Matsumoto. Neuronal Correlates of Goal-Based Motor Selection in the Prefrontal Cortex. *Science*, 301(5630):229–232, July 2003.

[85] William H Alexander and Joshua W Brown. Medial prefrontal cortex as an action-outcome predictor. *Nature Neuroscience*, 14(10):1338–1344, September 2011.

[86] Massimo Silvetti, William Alexander, Tom Verguts, and Joshua W Brown. From conflict management to reward-based decision making: Actors and critics in primate medial frontal cortex. *Neuroscience and Biobehavioral Reviews*, pages 1–15, June 2013.

[87] Erie D Boorman, Matthew F Rushworth, and Tim E Behrens. Ventromedial prefrontal and anterior cingulate cortex adopt choice and default reference frames during sequential multi-alternative choice. *Journal of Neuroscience*, 33(6):2242–2253, 2013.

[88] Paul W Burgess, Iroise Dumontheil, and Sam J Gilbert. The gateway hypothesis of rostral prefrontal cortex (area 10) function. *Trends in Cognitive Sciences*, 11 (7):290–298, 2007.

[89] Uta Frith and Christopher D Frith. Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 358 (1431):459–473, March 2003.

[90] Etienne Koechlin and Alexandre Hyafil. Anterior Prefrontal Function and the Limits of Human Decision-Making. *Science*, 318:594–598, October 2007.

[91] Etienne Koechlin. The role of the anteriorprefrontal cortex in human cognition. *Nature*, pages 1–4, May 1999.

[92] Nathaniel D Daw, John P O'Doherty, Peter Dayan, Ben Seymour, and Raymond J Dolan. Cortical substrates for exploratory decisions in humans. *Nature*, 441(7095): 876–879, June 2006.

[93] Erie D Boorman, Timothy E J Behrens, Mark W Woolrich, and Matthew F S Rushworth. How Green Is the Grass on the Other Side? Frontopolar Cortex and the Evidence in Favor of Alternative Courses of Action. *Neuron*, 62(5):733–743, June 2009.

[94] Mael Donoso, Anne G E Collins, and Etienne Koechlin. Foundations of human reasoning in the prefrontal cortex. *Science*, 344(6191):1481–1486, June 2014.

[95] Stephen M Fleming, Rimona S Weil, Zoltan Nagy, Raymond J Dolan, and Geraint Rees. Relating introspective accuracy to individual differences in brain structure. *Science*, 329(5998):1541–1543, September 2010.

[96] Stephen M Fleming and Raymond J Dolan. The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594):1338–1349, April 2012.

[97] Kent C Berridge and Terry E Robinson. Parsing reward. *Trends in Neurosciences*, 26(9):507–513, September 2003.

[98] Daniel Kahneman and Amos Tversky. Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2):263–292, March 1979.

[99] Amos Tversky and Richard H Thaler. Anomalies: Preference Reversals. *The Journal of Economic Perspectives*, 4(2):201–211, June 1990.

[100] Amos Tversky and Daniel Kahneman. The Framing of Decisions and the Psychology of Choice. *Science*, 211(4481):453–458, June 1981.

[101] Leon Festinger. *A theory of cognitive dissonance*, volume 2. Stanford university press, 1962.

[102] Keise Izuma and Kenji Matsumoto. Neural correlates of cognitive dissonance and choice-induced preference change. *Proceedings of the National Academy of Sciences*, pages 1–6, December 2010.

[103] Moti Salti, Imen El Karoui, Mathurin Maillet, and Lionel Naccache. Cognitive Dissonance Resolution Is Related to Episodic Memory. *PLoS ONE*, 9(9):e108579–9, September 2014.

[104] R Duncan Luce. The Choice Axiom after Twenty Years. *Journal of Mathematical Psychology*, 15:215–233, December 1977.

[105] Madeleine E Sharp, Jayalakshmi Viswanathan, Linda J Lanyon, and Jason J S Barton. Sensitivity and Bias in Decision-Making under Risk: Evaluating the Perception of Reward, Its Probability and Value. *PLoS ONE*, 7(4):e33460, April 2012.

[106] Sabrina M Tom, Craig R Fox, Christopher Trepel, and Russell A Poldrack. The neural basis of loss aversion in decision-making under risk. *Science*, 315(5811): 515–518, 2007.

[107] Christopher M Bishop. *Pattern Recognition and Machine Learning*, volume 4. Springer, August 2006.

[108] Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT PRESS, June 1998.

[109] Robert A Rescorla and Allan R Wagner. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning current research and theory*, 1972.

[110] Nathaniel D Daw, Yael Niv, and Peter Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8(12):1704–1711, November 2005.

[111] Amir Dezfouli and Bernard W Balleine. Actions, Action Sequences and Habits: Evidence That Goal-Directed and Habitual Action Control Are Hierarchically Organized. *PLoS Computational Biology*, 9(12):e1003364–14, December 2013.

[112] Kenji Doya, Kazuyuki Samejima, Ken-Ichi Katagiri, and Mitsuo Kawato. Multiple model-based reinforcement learning. *Neural computation*, 14(6):1347–1369, 2002.

[113] Kazuyuki Samejima and Kenji Doya. Multiple Representations of Belief States and Action Values in Corticobasal Ganglia Loops. *Annals of the New York Academy of Sciences*, 1104(1):213–228, April 2007.

[114] Jan Glascher, Nathaniel Daw, Peter Dayan, and John P O Doherty. States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning. *Neuron*, 66(4):585–595, May 2010.

[115] Nathaniel D Daw, Samuel J Gershman, Ben Seymour, Peter Dayan, and Raymond J Dolan. Model-Based Influences on Humans' Choices and Striatal Prediction Errors. *Neuron*, 69(6):1204–1215, March 2011.

[116] Suzanne N Haber. The primate basal ganglia: parallel and integrative networks. *Journal of Chemical Neuroanatomy*, 26(4):317–330, December 2003.

[117] Wolfram Schultz, Peter Dayan, and P Read Montague. A Neural Substrate of Prediction and Reward. *Science*, 275:1593–1599, March 1997.

[118] Wolfram Schultz. Predictive Reward Signal of Dopamine Neurons. *Journal of Neurophysiology*, 80:1–28, July 1998.

[119] Christopher D Fiorillo, Philippe N Tobler, and Wolfram Schultz. Discrete Coding of Reward Probability and Uncertainty by Dopamine Neurons. *Science*, 299(5614): 1898–1902, March 2003.

[120] Michael J Frank, Lauren C Seeberger, and Randall C OReilly. By Carrot or by Stick: Cognitive Reinforcement Learning in Parkinsonism . *Science*, 306(1): 1940–1943, December 2004.

[121] Stefano Palminteri, Mael Lebreton, Yulia Worbe, David Grabli, Andreas Hartmann, and Mathias Pessiglione. Pharmacological modulation of subliminal learning in Parkinson's and Tourette's syndromes. *Proceedings of the National Academy of Sciences*, 106(45):19179–19184, November 2009.

[122] Guillaume Sescousse, Jérôme Redouté, and Jean-Claude Dreher. The architecture of reward value coding in the human orbitofrontal cortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 30(39):13095–13104, September 2010.

[123] John P O'Doherty, Peter Dayan, Johannes Schultz, Ralf Deichmann, Karl Friston, and Raymond J Dolan. Dissociable Roles of Ventral and Dorsal Striatum in Instrumental Conditioning. *Science*, 304(5669):452–454, April 2004.

[124] John P O'Doherty, Alan Hampton, and Hackjin Kim. Model-Based fMRI and Its Application to Reward Learning and Decision Making. *Annals of the New York Academy of Sciences*, 1104(1):35–53, April 2007.

[125] Mathias Pessiglione, Ben Seymour, Guillaume Flandin, Raymond J Dolan, and Chris D Frith. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, 442(7106):1042–1045, August 2006.

[126] Léon Tremblay and Wolfram Schultz. Relative reward preference in primate orbitofrontal cortex. *Nature*, 398(6729):704–708, 1999.

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

[128] Helen C Barron, Raymond J Dolan, and Timothy E J Behrens. Online evaluation of novel choices by simultaneous representation of multiple memories. *Nature Publishing Group*, 16(10):1492–1498, September 2013.

[129] Sébastien Bouret and Barry J Richmond. Ventromedial and Orbital Prefrontal Neurons Differentially Encode Internally and Externally Driven Motivational Values in Monkeys. *Journal of Neuroscience*, 30(25):8591–8601, June 2010.

[130] Kenji Matsumoto, Wataru Suzuki, and Keiji Tanaka. Neuronal correlates of goal-based motor selection in the prefrontal cortex. *Science*, 301(5630):229–232, July 2003.

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

[132] Peter H Rudebeck, Mark J Buckley, Mark E Walton, and Matthew F S Rushworth. A role for the macaque anterior cingulate gyrus in social valuation. *Science*, 313 (5791):1310–1312, 2006.

[133] Timothy E J Behrens, Laurence T Hunt, Mark W Woolrich, and Matthew F S Rushworth. Associative learning of social value. *Nature*, 456(7219):245–249, November 2008.

[134] Xinying Cai and Camillo Padoa-Schioppa. Neuronal Encoding of Subjective Value in Dorsal and Ventral Anterior Cingulate Cortex. *Journal of Neuroscience*, 32(11): 3791–3808, March 2012.

[135] Nathalie Camille, Ami Tsuchida, and Lesley K Fellows. Double dissociation of stimulus-value and action-value learning in humans with orbitofrontal or anterior cingulate cortex damage. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 31(42):15048–15052, October 2011.

[136] Sean B Ostlund and Bernard W Balleine. Orbitofrontal cortex mediates outcome encoding in Pavlovian but not instrumental conditioning. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 27(18):4819–4825, May 2007.

[137] Peter H Rudebeck, Timothy E Behrens, Steven W Kennerley, Mark G Baxter, Mark J Buckley, Mark E Walton, and Matthew F S Rushworth. Frontal Cortex Subregions Play Distinct Roles in Choices between Actions and Stimuli. *Journal of Neuroscience*, 28(51):13775–13785, December 2008.

[138] Roland Peyron, Bernard Laurent, and Luis Garcia-Larrea. Functional imaging of brain responses to pain. A review and meta-analysis (2000). *Neurophysiol Clin*, 30:263–288, December 2000.

[139] Stefano Palminteri, Damian Justo, Céline Jauffret, Beth Pavlicek, Aurélie Dauta, Christine Delmaire, Virginie Czernecki, Carine Karachi, Laurent Capelle, Alexandra Durr, and Mathias Pessiglione. Critical Roles for Anterior Insula and Dorsal Striatum in Punishment-Based Avoidance Learning. *Neuron*, 76(5):998–1009, December 2012.

[140] Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and N D Goodman. How to Grow a Mind: Statistics, Structure, and Abstraction. *Science*, 331:1279–1285, March 2011.

[141] Mike Oaksford and Nick Chater. Précis of Bayesian Rationality: The Probabilistic Approach to Human Reasoning. *Behavioral and Brain Sciences*, 32(01):69–84, February 2009.

[142] Alexandre Pouget, Jeffrey M Beck, Wei Ji Ma, and Peter E Latham. Probabilistic brains: knowns and unknowns. *Nature Neuroscience*, 16(9):1170–1178, August 2013.

[143] Angela J Yu and Peter Dayan. Uncertainty, Neuromodulation, and Attention. *Neuron*, 46(4):681–692, May 2005.

[144] Elise Payzan-LeNestour and Peter Bossaerts. Risk, Unexpected Uncertainty, and Estimation Uncertainty: Bayesian Learning in Unstable Settings. *PLoS Computational Biology*, 7(1):e1001048, January 2011.

[145] Jonathan D Cohen and Angela J Yu. Sequential effects: superstition or rational behavior? *Advances in neural information processing systems*, pages 1873–1880, 2009.

[146] Thomas L Griffiths, Nick Chater, Charles Kemp, Amy Perfors, and Joshua B Tenenbaum. Probabilistic models of cognition: exploring representations and inductive biases. *Trends in Cognitive Sciences*, 14(8):357–364, August 2010.

[147] Sharon Goldwater, Thomas L Griffiths, and Mark Johnson. A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54, July 2009.

[148] Charles Kemp and Joshua B Tenenbaum. Structured statistical models of inductive reasoning. *Psychological Review*, 116(1):20–58, 2009.

[149] Fei Xu and Joshua B Tenenbaum. Word learning as Bayesian inference. *Psychological Review*, 114(2):245–272, 2007.

[150] Karl Friston. The anatomy of choice: active inference and agency. *Frontiers in human neuroscience*, 7(598):1–18, September 2013.

[151] Karl Friston, James Kilner, and Lee Harrison. A free energy principle for the brain. *Journal of Physiology-Paris*, 100(1-3):70–87, July 2006.

[152] Karl J Friston, Jean Daunizeau, James Kilner, and Stefan J Kiebel. Action and behavior: a free-energy formulation. *Biological Cybernetics*, 102(3):227–260, February 2010.

[153] Yael Niv, Reka Daniel, Andra Geana, Samuel J Gershman, Yuan Chang Leong, Angela Radulescu, and Robert C Wilson. Reinforcement Learning in Multidimensional Environments Relies on Attention Mechanisms. *Journal of Neuroscience*, 35(21):8145–8157, May 2015.

[154] Andra Geana and Yael Niv. Causal Model Comparison Shows That Human Representation Learning Is Not Bayesian. *Cold Spring Harbor Symposia on Quantitative Biology*, pages 024851–9, May 2015.

[155] Miguel P Eckstein, Craig K Abbey, Binh T Pham, and Steven S Shimozaki. Perceptual learning through optimization of attentional weighting: Human versus optimal Bayesian learner. *Journal of Vision*, 4(12):3–3, December 2004.

[156] Jill X O'Reilly, Saad Jbabdi, and Timothy EJ Behrens. How can a Bayesian approach inform neuroscience? *European Journal of Neuroscience*, 35(7):1169–1179, 2012.

[157] Giorgio Coricelli, Hugo D Critchley, Mateus Joffily, John P O'Doherty, Angela Sirigu, and Raymond J Dolan. Regret and its avoidance: a neuroimaging study of choice behavior. *Nature Neuroscience*, 8(9):1255–1262, August 2005.

[158] Giorgio Coricelli and Aldo Rustichini. Counterfactual thinking and emotions: regret and envy learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1538):241–247, December 2009.

[159] Stefano Palminteri, Mehdi Khamassi, Mateus Joffily, and Giorgio Coricelli. Contextual modulation of value signals in reward and punishment learning. *Nature Communications*, 6:1–14, August 2015.

[160] Anne Collins and Etienne Koechlin. Reasoning, Learning, and Creativity: Frontal Lobe Function and Human Decision-Making. *PLoS Biology*, 10(3):e1001293–16, March 2012.

[161] Gregory Schraw. Promoting general metacognitive awareness. *Instructional Science*, 26:113–125, December 1998.

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

[163] George Bush, Phan Luu, and Michael I Posner. Cognitive and emotional influences in anterior cingulate cortex. *Trends in Cognitive Sciences*, 4(6):215–222, 2000.

[164] Amit Etkin, Tobias Egner, and Raffael Kalisch. Emotional processing in anterior cingulate and medial prefrontal cortex. *Trends in Cognitive Sciences*, 15(2):85–93, February 2011.

[165] Alexander J Shackman, Tim V Salomons, Heleen A Slagter, Andrew S Fox, Jameel J Winter, and Richard J Davidson. The integration of negative affect, pain and cognitive control in the cingulate cortex. *Nature Publishing Group*, 12 (3):154–167, March 2011.

[166] Mehdi Keramati, Amir Dezfouli, and Payam Piray. Speed/Accuracy Trade-Off between the Habitual and the Goal-Directed Processes. *PLoS Computational Biology*, 7(5):e1002055–21, May 2011.

[167] David H Brainard. The Psychophysics Toolbox. *Spatial Vision*, 10(4):433–436, December 1997.

[168] Douglas M Hawkins. The Problem of Overfitting. *Journal of Chemical Information and Modeling*, 44(1):1–12, January 2004.

[169] Wolfgang Gaissmaier and Lael J Schooler. The smart potential behind probability matching. *Cognition*, 109(3):416–422, December 2008.

[170] Daniel Kahneman and Amos Tversky. Choices, Values and Frames. *American Psychologist*, 39(4):341–350, April 1984.

[171] Amos Tversky and Daniel Kahneman. Judgment under Uncertainty: Heuristics and Biases. *Science*, 185:1125–1131, September 1974.

[172] Hang Zhang and Laurence T Maloney. Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in neuroscience*, 6(1):1–14, January 2012.

[173] Mehdi Khamassi, Pierre Enel, Peter Ford Dominey, and Emmanuel Procyk. Medial prefrontal cortex and the adaptive regulation of reinforcement learning parameters. *Prog Brain Res*, 202:441–464, 2013.

[174] Ralf Deichmann, Jay A Gottfried, Chloe A Hutton, and Robert Turner. Optimized EPI for fMRI studies of the orbitofrontal cortex. *NeuroImage*, 19(2):430–441, June 2003.

[175] John O Doherty, Alan N Hampton, and Hackjin Kim. Model-Based fMRI and Its Application to Reward Learning and Decision Making. *Annals of the New York Academy of Sciences*, 1104(1):35–53, April 2007.

[176] Jack Grinband, Tor D Wager, Martin Lindquist, Vincent P Ferrera, and Joy Hirsch. Detection of time-varying signals in event-related fMRI designs. *NeuroImage*, 43(3):509–520, November 2008.

[177] Joseph F Hair, William C Black, Barry J Babin, Rolph E Anderson, and Ronald L Tatham. *Multivariate data analysis*, volume 6. Pearson Prentice Hall Upper Saddle River, NJ, 2006.

[178] Nikolaus Kriegeskorte, W Kyle Simmons, Patrick S F Bellgowan, and Chris I Baker. Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience*, 12(5):535–540, April 2009.

[179] Guillaume Sescousse, Xavier Caldú, Bàrbara Segura, and Jean-Claude Dreher. Processing of primary and secondary rewards: A quantitative meta-analysis and review of human functional neuroimaging studies. *Neuroscience and Biobehavioral Reviews*, 37(4):681–696, May 2013.

[180] Julia Trommershäuser, Laurence T Maloney, and Michael S Landy. Decision making, movement planning and statistical decision theory. *Trends in Cognitive Sciences*, 12(8):291–297, August 2008.

[181] Stanislas Dehaene. Psychologie cognitive expérimentale. In Editions Fayard, editor, *Vers une science de la vie mentale*, pages 277–301. November 2006.

[182] Mael Lebreton, Maxime Bertoux, Claire Boutet, Stephane Lehericy, Bruno Dubois, Philippe Fossati, and Mathias Pessiglione. A Critical Role for the Hippocampus in the Valuation of Imagined Outcomes. *PLoS Biology*, 11(10):e1001684–13, October 2013.

[183] Steven W Kennerley, Mark E Walton, Timothy EJ Behrens, Mark J Buckley, and Matthew FS Rushworth. Optimal decision making and the anterior cingulate cortex. *Nature Neuroscience*, 9(7):940–947, 2006.

[184] Jeffrey M Beck, Wei Ji Ma, Xaq Pitkow, Peter E Latham, and Alexandre Pouget. Not Noisy, Just Wrong: The Role of Suboptimal Inference in Behavioral Variability. *Neuron*, 74(1):30–39, April 2012.

[185] Ralph Hertwig, Greg Barron, Elke U Weber, and Ido Erev. Decisions From Experience and the Effect of Rare Events in Risky Choice. *Psychological Science*, 15 (8):534–539, October 2004.

[186] Philippe N Tobler, George I Christopoulos, John P O'Doherty, Raymond J Dolan, and Wolfram Schultz. Neuronal distortions of reward probability without choice. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 28(45):11703–11711, November 2008.

[187] Rosanna Michalczuk, Henrietta Bowden-Jones, Antonio Verdejo-Garcia, and Luke Clark. Impulsivity and cognitive distortions in pathological gamblers attending the UK National Problem Gambling Clinic: a preliminary report. *Psychological Medicine*, 41(12):2625–2635, June 2011.

[188] Christopher Summerfield and Konstantinos Tsetsos. Do humans make good decisions? *Trends in Cognitive Sciences*, 19(1):27–34, January 2015.

[189] Falk Lieder, Ming Hsu, and Thomas L Griffiths. The high availability of extreme events serves resource-rational decision-making. In *Proceedings of the th Annual*

Conference of the Cognitive Science Society. Austin, TX Cognitive Science Society, pages 1–6, 2014.

[190] Christopher Summerfield, Timothy E Behrens, and Etienne Koechlin. Perceptual Classification in a Rapidly Changing Environment. *Neuron*, 71(4):725–736, August 2011.

[191] Bradley B Doll, Dylan A Simon, and Nathaniel D Daw. The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology*, 22(6):1075–1081, December 2012.

[192] Amir Dezfouli and Bernard W Balleine. Actions, Action Sequences and Habits: Evidence That Goal-Directed and Habitual Action Control Are Hierarchically Organized. *PLoS Computational Biology*, 9(12):e1003364–14, December 2013.

[193] Kate Teffer and Katerina Semendeferi. Human prefrontal cortex: Evolution, development, and pathology. In Michael A Hofman and Dean Falk, editors, *Progress in brain research*, page 191. 2012.

[194] Daniel Kersten and Alan Yuille. Bayesian models of object perception. *Current Opinion in Neurobiology*, 13(2):150–158, April 2003.

[195] Konstantinos Tsetsos, Valentin Wyart, S Paul Shorkey, and Christopher Summerfield. Neural mechanisms of economic commitment in the human medial prefrontal cortex. *eLife*, 3:1–17, October 2014.

[196] Fabrizio Esposito, Alessandro Bertolino, Tommaso Scarabino, Valeria Latorre, Giuseppe Blasi, Teresa Popolizio, Gioacchino Tedeschi, Sossio Cirillo, Rainer Goebel, and Francesco Di Salle. Independent component model of the default-mode brain function: Assessing the impact of active thinking. *Brain research bulletin*, 70(4):263–269, 2006.

[197] Peter Fransson. How default is the default mode of brain function? *Neuropsychologia*, 44(14):2836–2845, January 2006.

[198] Debra A Gusnard, Erbil Akbudak, Gordon L Shulman, and Marcus E Raichle. Medial prefrontal cortex and self-referential mental activity: Relation to a default mode of brain function. *Proceedings of the National Academy of Sciences of the United States of America*, 98(7):4259–4264, March 2001.

[199] John Clithero and Antonio Rangel. Informatic parcellation of the network involved in the computation of subjective value. *Social Cognitive and Affective Neuroscience*, pages 1–53, April 2013.

[200] Miriam Cornelia Klein-Flugge, Helen C Barron, Kay H Brodersen, Raymond J Dolan, and Timothy E J Behrens. Segregated Encoding of Reward-Identity and Stimulus-Reward Associations in Human Orbitofrontal Cortex. *Journal of Neuroscience*, 33(7):3202–3211, February 2013.

[201] Ryan K Jessup and John P O'Doherty. Distinguishing informational from value-related encoding of rewarding and punishing outcomes in the human brain. *European Journal of Neuroscience*, 39(11):2014–2026, May 2014.

[202] Thorsten Kahnt, Soyoung Q Park, John-Dylan Haynes, and Philippe N Tobler. Disentangling neural representations of value and salience in the human brain. *Proceedings of the National Academy of Sciences*, 111(13):5000–5005, April 2014.

[203] Benedetto De Martino, Stephen M Fleming, Neil Garrett, and Raymond J Dolan. Confidence in value-based choice. *Nature Neuroscience*, 16(1):105–110, December 2012.

[204] Stephen M Fleming, Brian Maniscalco, Yoshiaki Ko, Namema Amendi, Tony Ro, and Hakwan Lau. Action-specific disruption of perceptual confidence. *Psychological Science*, 26(1):89–98, January 2015.

[205] Mael Lebreton, Raphaëlle Abitbol, Jean Daunizeau, and Mathias Pessiglione. Automatic integration of confidence in the brain valuation signal. *Nature Neuroscience*, 18(8):1159–1167, July 2015.

List of Figures

List of Tables