Mechanisms of conscious and unconscious interpretative processes

Imen El Karoui

Université Pierre et Marie Curie

Ecole doctorale Cerveau Cognition Comportement

Institut du Cerveau et de la Moelle épinière

Physiological investigation of clinically normal and impaired cognition

Mechanisms of conscious and unconscious interpretative processes

Par Imen El Karoui

Thèse de doctorat de Neurosciences cognitives

Dirigée par Lionel Naccache

Présentée et soutenue publiquement le 8 Avril 2015

Devant un jury composé de :

Eva VAN DEN BUSSCHE	MCU	Vrije Universiteit Brussel	Rapporteuse
Keise IZUMA	MCU	University of York	Rapporteur
Catherine TALLON-BAUDRY	DR	Ecole Normale Supérieure	Examinatrice
Victor LAMME	PU	University of Amsterdam	Examinateur
Philippe FAURE	DR	Université Pierre et Marie Curie	Examinateur
Lionel NACCACHE	PU-PH	Université Pierre et Marie Curie	Directeur

ACKNOWLEDGEMENTS

Je souhaiterais tout d'abord remercier Lionel Naccache, mon directeur de thèse. Merci pour ton enthousiasme contagieux et le foisonnement d'idées qui ressortaient de chacune de nos discussions, malgré mon côté rabat-joie statistique. C'était un grand plaisir de travailler à tes côtés pendant ces quatre années. J'ai appris énormément tant du point de vue scientifique que du point de vue humain. Merci notamment pour toutes les discussions que l'on a pu avoir sur les patients que tu voyais en consultation ou dans le cadre des recherches sur les désordres de conscience, qui m'ont permis de mieux appréhender les aspects cliniques liés à nos recherches.

I would also like to thank all the members of the jury: Eva Van Den Bussche, Keise Izuma, Catherine Tallon-Baudry, Victor Lamme and Philippe Faure. It is a great honor to be able to discuss my work with some of the people who inspired it.

Je voudrais ensuite remercier les membres de l'équipe dans laquelle j'ai travaillé et sans qui ce travail n'aurait jamais été possible : merci, thank you, gracias (with Argentinian and Spanish accents ☺), grazie, danke, dankjewel, obrigado, شكراً, متشكرم, הדות et ευχαριστώ !! It was a great experience working with you all and I really appreciated the cultural diversity of the lab (and I'm not talking only about food!). Je remercie tout particulièrement Laurent Cohen pour toutes ces remarques constructives sur mon travail mais aussi son humour et son auto-dérision. Special thanks also to Moti Salti, I loved our passionate meetings and all your stories about Eyal! Un grand merci aussi à Mariam Chammat pour son enthousiasme à propos des biais cognitifs et de tout le reste, sa gentillesse et ses multiples relectures de ce manuscrit. Merci aussi à Jean-Rémi King d'avoir partagé tant de moments au labo et de m'avoir donné des nouvelles de son vélo, Jacobo Sitt for constant support through these four years, Benjamin Rohaut et ses histoires de réa, Camille Rozier et ses délicieux cupcakes, Fabien Vinckier et son rire communicatif, Pierre Bourdillon et ses flops, Claire Sergent pour le livre des gâteaux de mamie, Mariana Babo-Rebelo pour nos pauses café et Simon van Gaal for dancing on tables. Merci aussi à : Sami Abboud, Florence Bouhali, Athena Demertzi, Denis Engemann, Diane

Lazard, Diana Lopez-Barroso, Mathurin Maillet, Valeria Mongelli, Marion Quirins, Federico Raimundo, Tal Seidel.

Je souhaiterais aussi remercier mes différents collaborateurs. Tout d'abord, merci à Claude Adam, Vincent Navarro et Dominique Hasboun de m'avoir permis de travailler avec les patients épileptiques implantés et de m'avoir aidé dans la localisation de leurs électrodes. Je remercie aussi tout particulièrement Joseph et Corinne, sans qui l'enregistrement de ces patients n'aurait pas été possible. Je tiens à remercier Karim N'Diaye pour avoir lancé le mouvement de coordination entre les différentes équipes travaillant avec les patients épileptiques, ainsi que Katia Lehongre pour la poursuite de cette coordination. Je remercie aussi Sara Fernandez-Vidal et Fernando Pérez Garcia pour le développement d'outils de localisation automatique. Je profite aussi de cet espace pour remercier Paolo Bartolomeo, Dimitri Bayle et Alexia Bourgeois, ainsi que Nathalie George, Vera Dinckelaker et Josefien Huijgen et enfin Stanislas Dehaene pour nos différents projets avec les patients épileptiques. Merci ensuite à Carine Karachi, Alister Rogers et Brian Lau pour m'avoir fait découvrir l'intérieur d'un bloc opératoire de neurochirurgie. Enfin, merci au CENIR pour leur aide dans le cadre de notre manip IRM.

Au sein de l'ICM, je tiens à remercier l'équipe MBB pour m'avoir acceptée dans vos soirées et en particulier Alizée pour tes conseils IRM et Marie pour nos déjeuners. Je remercie aussi les Ajités pour m'avoir permis de faire la plupart de mes réunions sur un canapé, ainsi que l'équipe du théâtre pour m'avoir habituée à la scène de l'auditorium.

Je remercie aussi tous les sujets qui ont participé à mes expériences, et en particulier les patients qui ont accepté de passer une partie de leur temps devant mon ordinateur pendant leur hospitalisation.

Je remercie ensuite le Fonds AXA pour la Recherche qui a financé cette thèse et tout particulièrement Raphaëlle Lalo et Isabelle Delaporte pour leur accueil au cours des Pop'Days.

Un grand merci aussi à tous mes amis pour tous les moments passés ensemble, autour de déjeuners à côté du muséum, de week ends minettes, de cours de gym su, d'apéros du dimanche ou d'un brunch déguisé…

Merci aussi à toute ma famille, élargie officiellement récemment, et notamment à mes parents pour leur soutien permanent.

Enfin, merci à Florent pour tous les moments que nous partageons.

ABSTRACT

When we perceive a word, a picture or a sound, we do not access an 'objective' representation of them. Rather we gain immediate access to a meaningful mental content colored with subjective belief. This interpretation reflects the combination of our prior knowledge about the world with data sampled in the environment. An interesting issue is to understand how we deal with inconsistencies between our prior knowledge and the data from the environment. During this PhD, responses to inconsistencies both in the environment and in subjects' own behavior were explored.

The first series of studies address how subjects process regularities in the environment, as well as stimuli contradicting these regularities, to guide their behavior. A second aim of these studies was to understand how these processes relate to conscious access by testing whether subjects needed to be conscious of the stimuli involved in the regularities, and whether they were conscious of the regularities themselves. To do so, two levels of auditory regularities were studied in epileptic patients implanted with intracranial electrodes. This setting allowed exploring the neural signatures of these regularities with high spatial and temporal resolution. In the second experiment, we used a paradigm derived from the Stroop task to test how subjects could change their priors when responding to frequent conscious or unconscious response conflicts. Behavioral measures and scalp EEG were used to assess changes in subjects' strategy when processing trials conflicting with current expectations.

In the second series of studies, we analyzed how subjects adapt their interpretations when confronted with inconsistencies in their own behavior, using the framework of cognitive dissonance. Mechanisms of cognitive dissonance reduction following difficult choices were explored to assess whether this effect rely on reportable information or not. Notably, the implication of explicit memory was tested in a behavioral experiment and in an fMRI study.

The results of these four studies are discussed around two main issues. First, these results highlight the existence of processes which rely on conscious stimuli but are not conscious themselves. Second, we examine what could explain our tendency to constantly seek consistency both in the external world and in our own behavior.

Keywords: interpretations, consciousness, cognitive control, cognitive dissonance

RESUME

Lorsqu'une représentation accède à la conscience, ce n'est pas simplement une représentation « objective » d'un mot, d'une image ou d'un son, mais plutôt une interprétation subjective, c'est-à-dire un contenu mental, sélectionné parmi plusieurs possibles et associé à un certain degré de croyance subjective. Cette interprétation reflète la combinaison de nos connaissances sur le monde avec les données de notre environnement. Il est donc intéressant de comprendre comment ces interprétations sont mises à jour lorsque l'on est confronté à des informations qui ne sont pas cohérentes avec nos connaissances sur le monde. Dans cette thèse, nous nous sommes intéressés aux incohérences existant dans l'environnement, mais aussi dans le comportement des individus.

Dans une première série d'études, nous avons étudié l'apprentissage de régularités dans l'environnement et la réponse cérébrale aux stimuli ne correspondant pas à ces régularités. Nous avons aussi analysé les relations entre ces processus et la conscience d'accès, afin de déterminer si les sujets devaient nécessairement prendre conscience des stimuli pour pouvoir identifier les régularités, mais aussi s'ils étaient conscients de ces régularités elles-mêmes. Pour ce faire, nous avons utilisé un paradigme expérimental incluant deux niveaux de régularités dans un flux sonore. Nous avons enregistré les réponses cérébrales à ces différents stimuli chez des patients épileptiques ayant des électrodes intracérébrales dans le cadre d'un bilan pré-chirurgical. Ces enregistrements bénéficient d'une bonne résolution spatiale et temporelle, ce qui nous a permis de comprendre de façon détaillée les réponses cérébrales associées à ces régularités. Dans une seconde étude, nous avons développé un paradigme expérimental à partir de la tâche de Stroop pour tester les réponses à de fréquents conflits, perçus consciemment ou non. Nous avons utilisé des mesures comportementales ainsi que l'électroencéphalographie (EEG) pour déterminer quelle stratégie était utilisée par les sujets lors des essais en contradiction avec leurs attentes.

Dans une seconde série d'études, nous avons étudié comment les sujets traitent les incohérences dans leur propre comportement, dans le cadre de la théorie de la dissonance cognitive. Nous avons utilisé le paradigme du choix libre, afin d'analyser la réduction de la dissonance liée à la réalisation de choix difficiles. Nous avons identifié un rôle crucial de la mémoire dans cet effet. Ce rôle a été exploré grâce à une étude comportementale et à une étude en IRM fonctionnelle.

Les résultats de ces quatre études sont discutés dans ce manuscrit autour de deux questions clés. Tout d'abord, ces résultats mettent en évidence l'existence de processus qui nécessitent que les stimuli sur lesquels ils opèrent soient conscients, mais qui ne sont pas conscients eux-mêmes. Ensuite, nous apportons quelques éléments de discussion permettant de comprendre pourquoi l'on tend à chercher de la cohérence, aussi bien dans notre environnement que dans notre comportement.

Mots-clés : interprétations, conscience, contrôle cognitif, dissonance cognitive

LIST OF PUBLICATIONS

Articles presented in this dissertation

- **Chapter 2:** El Karoui I., King J.-R., Sitt J., Meyniel F., Van Gaal S., Hasboun D., Adam C., Navarro V., Baulac M., Dehaene S., Cohen L., Naccache L. (2014). Event-Related Potential, Time-frequency, and Functional Connectivity Facets of Local and Global Auditory Novelty Processing: An Intracranial Study in Humans. *Cerebral Cortex,* pii: bhu-143
- **Chapter 3:** El Karoui I., Christoforidis K., Naccache L. (under review). Do acquisition and transfer of strategy require conscious access?
- **Chapter 4:** Salti, M.*, El Karoui, I.*, Maillet, M., & Naccache, L. (2014). Cognitive Dissonance Resolution Is Related to Episodic Memory. *PloS One*, 9(9), 1–8. (* these authors contributed equally to the work)

Other publications

- King, J. R., Faugeras, F., Gramfort, A., Schurger, A., El Karoui, I., Sitt, J. D., Rohaut, B., Wacongne, C., Labyt, E., Bekinschtein, T., Cohen, L., Naccache, L., Dehaene, S. (2013). Single-trial decoding of auditory novelty responses facilitates the detection of residual consciousness. *NeuroImage,* 83, 726–738.
- King, J.-R.*, Sitt, J. D.*, Faugeras, F., Rohaut, B., El Karoui, I., Cohen, L., Naccache, L., Dehaene, S. (2013). Information sharing in the brain indexes consciousness in noncommunicative patients. *Current Biology*, 23(19), 1914–9.
- Naccache, L., Sportiche, S., Strauss, M., El Karoui, I., Sitt, J., & Cohen, L. (2014). Imaging "top-down" mobilization of visual information: a case study in a posterior split-brain patient. *Neuropsychologia*, 53, 94–103.
- Sitt, J. D.*, King, J.-R.*, El Karoui, I., Rohaut, B., Faugeras, F., Gramfort, A., Cohen, L., Sigman, M., Dehaene, S., Naccache, L. (2014). Large scale screening of neural signatures of consciousness in patients in a vegetative or minimally conscious state. *Brain*, 137, 2258–70.
- Rohaut, B., Faugeras, F., Chausson, N., King, J.-R., El Karoui, I., Cohen, L., & Naccache, L. (2014). Probing ERP correlates of verbal semantic processing in patients with impaired consciousness. *Neuropsychologia*, 66C, 279–292.
- Huijgen, J.*, Dinkelacker, V.*, Lachat, F., Yahia-Cherif, L., El Karoui, I., Lemaréchal, J.-D., Adam, C., Hugueville, L., George, N. (under review) Amygdala processing of social cues from faces: an intracerebral EEG study

- Bayle, D., Bourgeois A., Navarro, V., Adam, C., Chica, A. B., Lupiañez, J., El Karoui, I., Bartolomeo, P. (under review) Dynamics of human attention revealed by intracerebral recordings

CONTENTS

'One can't believe impossible things.'

'I daresay you haven't had much practice,' said the Queen. 'When I was your age I always did it for half-an-hour a day. Why, sometimes I've believed as many as six impossible things before breakfast.'

Lewis Carroll, *Through the Looking Glass*

FOREWORD

Look at the picture on the left of Fig. 1.1. You can clearly see that the square A is darker than the square B. However, when joining both squares with two vertical stripes of the same shade of gray (Fig. 1.1 right), it becomes apparent that squares A and B are actually exactly the same color. Interestingly, even when you know that both squares are the same color, you cannot prevent your seeing square A as darker. This visual illusion (Adelson and Pentland, 1996) reflects that perception is a construct based on the integration of prior knowledge about the world with inputs from the environment: the illusion is due to the fact that our brain automatically interprets the darker area on the left of the cylinder as a shadow. Based on previous experiences, if the light source is on the right as suggested by the shadow, the square B would actually be brighter in full light and is thus interpreted as brighter than the square A.

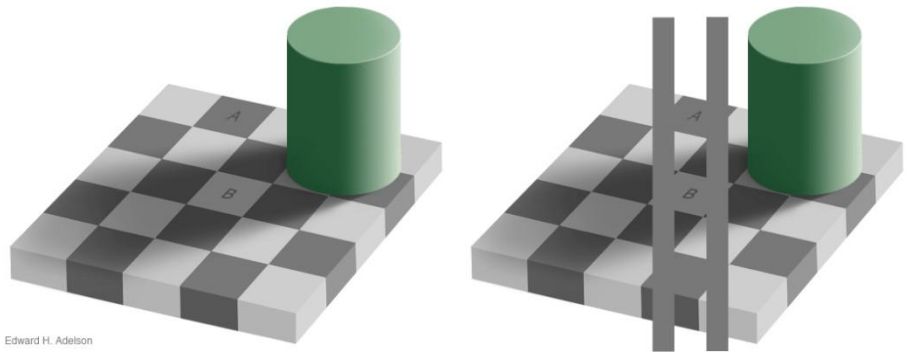

FIGURE 1.1: ADELSON'S CHECKERBOARD

In this example, the squares A and B are exactly the same color, but A is perceived as darker than B (adapted from Adelson and Pentland, 1996).

This integration of prior knowledge with inputs from the environment actually occurs whenever we perceive something. We gain immediate access to this integrated information that could be defined as a 'subjective interpretation', rather than to the 'objective' information provided by the environment. These interpretations can be related to low-level perceptual processes, as evidenced by Adelson's illusion, but also to higher level cognitive processes, as revealed by cognitive biases. One interesting difference in these types of interpretations is how much you can control them: in the case of low-level interpretations, you cannot control how the

information is integrated and you cannot help perceiving the square A as darker the square B, even when you know that they are the same color. On the other hand, higher level interpretations can be modified by new information that the subject is actively gathering. Therefore, there is an interesting interplay between subjective interpretations and conscious access.

In this thesis, I was particularly interested in how subjective interpretations maximize consistency and meaningfulness. In the introduction, I will provide several examples from neurological patients and control subjects highlighting this property. I will then present the framework within which this work was formulated. In Chapter 2 to 5, I will present the experiments I conducted, to better understand how subjects update their knowledge about the world when facing inconsistencies (both in the environment and in their own behavior) and how this process interacts with conscious access. Finally, I will discuss the results obtained in my experimental work in light of two main questions: what do we need to be conscious of? And what can explain our need of consistency?

CHAPTER 1. INTRODUCTION

1. Our interpretations are created to maximize consistency and meaningfulness

In this introduction, I will present a few examples which highlight the way our subjective interpretations tend to maximize the consistency between data from the environment and previous knowledge about the world. Interestingly, these interpretations are usually so meaningful that it is difficult to notice that they are actually constructed. Therefore, the examples presented in the following section are taken from the neuropsychology literature. Indeed, some patients create bizarre interpretations that can shed light on this process.

1.1. Evidence from neuropsychology

1.1.1. Split-brain patients

One of the most striking examples of our capacity to create coherent interpretations, to build stories about our perception and our behavior, can be found in the work of Sperry and Gazzaniga on split-brain patients. These patients suffered from epilepsy and had undergone surgery that severed their corpus callosum, creating a disconnection between their left and right hemispheres. The aim was to confine seizures to only one hemisphere and this surgery proved to be successful, with the elimination of almost all seizures, including unilateral ones (Gazzaniga, 1967). The operation did not result in major changes in patients' cognitive abilities. But closer observation revealed that the transmission of information between hemispheres was blocked, revealing differences between both hemispheres, and how they interact. In typical experimental settings, patients were asked to fixate the center of a screen and were shown different stimuli in their left and right visual hemifields (Gazzaniga et al., 1965). Interestingly, stimuli in the right visual hemifield (thus processed by the left hemisphere) could be described verbally by patients, but for stimuli in the left visual hemifield (processed by the right hemisphere), patients said they could not see anything. However, when asked to point to an object similar to the one presented in their left visual hemifield, they could do it easily with their left hand (Gazzaniga et al., 1965).

These results reflect the high hemispheric lateralization of language and the absence of information transfer between hemispheres after the callosotomy.

An interesting question is how the left hemisphere, the only one producing language, reacts to actions triggered by the right hemisphere. Ledoux and Gazzaniga developed an experiment to test this. Patients were presented with two different pictures, one in each hemifield, and then they were asked to pick images associated with these pictures, from a set of images displayed in

See "Finding False Memory" box in Gazzanigo (1998) – this figure is protected by copyright

FIGURE 1.2: EXAMPLE OF AN EXPERIMENT WITH A SPLIT-BRAIN PATIENT

Patient P.S is presented with one image in each hemifield and is asked to pick an item corresponding to each image with his left and right hands. When asked why he chose these items, he constructed a story combining both items, although his left hemisphere involved in language processing did not have access to the information presented in the left hemifield (adapted from Gazzaniga, 1998) .

front of them (Gazzaniga, 1989, 1998, 2000). For example, patient P.S. was presented with a snow scene picture in his left hemifield and a chicken claw in his right hemifield. Among the images presented in front of him, the one associated with the chicken claw was the chicken and the one associated with the snow scene was the snow shovel (Fig. 1.2). Subject P.S correctly picked these images with his right and left hands respectively. But when asked why he chose these items, he replied "Oh, that's simple. The chicken claw goes with the chicken and you need a shovel to clean out the chicken shed". Strikingly, the left hemisphere of this patient, observing the left hand response, interpreted it in the context of its own knowledge, which did not include the information of the snow scene, only presented to the right hemisphere. Note that in this case, the left hemisphere did not say "I don't know why I picked the shovel" but created an interpretation based on its available knowledge.

A similar experiment was conducted with patients showing language processing abilities both in the left and right hemispheres. These patients were presented with written action words in their left visual hemifield, processed by their right hemisphere (Gazzaniga et al., 1977). For example,

the word "laugh" was presented and the patient laughed, but when asked why, he did not refer to the command presented only to his right hemisphere but found an explanation: "You guys come up and test us every month. What a way to make a living!". Similarly, if the command "walk" was presented in the left hemifield, the patient would walk to the door and when asked where he is going, he would say "I'm going into the house to get a coke". In these examples again, the left hemisphere is interpreting the actions triggered by the right hemisphere in a coherent way, even though it does not have access to the relevant information.

These different studies in split-brain patients thus reveal our ability to attribute meaning to our own actions, even when no information about their causes is available.

1.1.2. Capgras syndrome

The Capgras syndrome also provides evidence of this ability of ours to create interpretations a posteriori. The first case was described by Capgras and Reboul-Lachaux in 1923. Mrs. M went to the police in June 1918 to warn them about the sequestration of thousands of people in the basement of her house and more generally in the basement of Paris. She was hospitalized in Sainte-Anne hospital and clinical examination revealed an interesting feature of her delusions: she was saying that many people around her, including her husband and daughters, were replaced by imposters (Capgras and Reboul-Lachaux, 1923). The Capgras syndrome was defined following this report: these patients are usually quite lucid in many aspects, but they see their close acquaintances as "imposters", i.e. reporting that they look like their close acquaintances, but they are not. When directly challenged, patients often justify their claims. For example, a patient explained that she could tell the difference between her real son and the imposter that has replaced him because he had "different colored eyes, was not as big and brawny, and that her real son would not kiss her" (Turner and Coltheart, 2010). Another example is given by Hirstein and Ramachandran (1997). Patient DS, when asked why he thought his father has been replaced by an impostor, replied "He looks exactly like my father, but he really isn't. He's a nice guy, but he isn't my father, Doctor'. Then the experimenter asked: "But why was this man pretending to be your father?" and DS replied: "That is what is so surprising,

Doctor – why should anyone want to pretend to be my father? Maybe my father employed him to take care of me – paid him some money so that he could pay my bills…"

A two-factor model was suggested to account for this syndrome (Langdon and Coltheart, 2000; Turner and Coltheart, 2010). A disconnection between face recognition and familiarity processes has been proposed as the first factor determining the content of delusions. This hypothesis relies on studies by Ellis et al (1997) and Hirstein and Ramachandran (1997). In these experiments, Capgras patients were presented with familiar and unfamiliar faces, while their skin conductance response (SCR), a marker of autonomic activity, was recorded. Capgras patients did not show any difference in SCR for familiar compared to unfamiliar faces, whereas in normal and psychiatric controls a significant difference was observed. Interestingly, however, they were able to recognize familiar faces (Ellis et al., 1997; Hirstein and Ramachandran, 1997). Capgras patients thus recognize the face of their close acquaintances but they do not have the usual emotional response associated with these faces. Interestingly, Tranel et al. (1995) reported the cases of four patients suffering from bilateral ventromedial frontal damage who presented a similar profile of SCRs: they were able to correctly recognize faces and to accurately attribute familiarity ratings, but they failed to generate discriminatory SCRs between familiar and unfamiliar faces. However, these patients did not develop a delusional belief about the identity of their acquaintances. These results suggest that a deficit in familiarity perception alone is not sufficient to trigger delusional beliefs. Nevertheless, note that the phenomenal experiences of patients with ventromedial prefrontal lesions have not been compared with those of Capgras patients and this similarity in SCR patterns might be related to different mechanisms, as it has been shown that patients with bilateral ventromedial prefrontal lesions fail to show normal SCR not only to faces but also to emotionally charged stimuli (Damasio et al., 1990).

Importantly, Capgras patients do not explain this failure to detect familiarity associated with the faces of close acquaintances by saying "I know this person is my husband but I no longer feel the warmth", instead they build a complex interpretation of this unusual feeling and say that "This

person is not my husband, he is an imposter". One potential explanation is that these patients present an additional lesion, possibly in the right prefrontal cortex (Alexander et al., 1979; Feinberg and Keenan, 2005) that prevents them from checking their interpretations. In an attempt to explain another deficit termed anosognosia, it has been argued that the left hemisphere aims at preserving consistency by explaining away any discrepancies, whereas the right hemisphere might provide a global monitoring and checking mechanisms on the interpretations produced by the left hemisphere (Ramachandran, 1995). A deficit in belief evaluation would explain why patients do not correct their interpretations, even though as reported by Young (1998), "if you ask [a Capgras patient] 'what would you think if I told you my wife had been replaced by an imposter', you will often get answers stating that it would be unbelievable, absurd, an indication that you had gone mad. Yet, these patients will claim that nonetheless this is exactly what has happened to their own relative" (Young, 1998).

Capgras patients thus provide coherent and meaningful, although bizarre, interpretations of their experience. The belief evaluation system is of particular interest to better understand mechanisms underlying the creation of interpretations. Its role has also been suggested to be important in another disorder, confabulations.

1.1.3. Confabulations

Indeed, some patients suffering from amnesia produce fictitious stories about events that never occurred, which are called "confabulations" (Schacter et al., 1998). They have first been described by Korsakoff in alcoholic patients with amnesia (Korsakoff, 1889, 1955). Metcalf et al. (2007) report the case of patient SD, a 36-year-old man, who began confabulating after a bicycle accident. When asked where he was, he would say in high-school, doing physical education, English and math. But at the time, he was actually in a rehabilitation center and had finished high-school twenty years before the interview. Interestingly, these patients often deny having a memory problem and when presented with information contradicting their beliefs, they would try to explain them, but fail to change their ideas about reality. For example, a patient convinced to be in Bordeaux when he was actually in Berne would agree that the view from his window

does not resemble Bordeaux, but would add "I am not crazy, I know that I am in Bordeaux!" (Schnider, 2003). It thus seems that for patients, confabulations are the honest narratives of their perceived reality, which is disturbed.

Several models have been proposed to explain this disorder. The first model posits that confabulations arise when the monitoring of ongoing reality fails (Schnider, 2003, 2013). Indeed an interesting feature of confabulations is that they can almost always be traced back to an actual event in the patient's past. Spontaneous confabulation thus seems to result from the inability to suppress evoked memories that do not belong to the ongoing reality. This process is mediated by the anterior limbic system and more specifically by the posterior orbitofrontal cortex, which is typically injured in patients producing spontaneous confabulations. Note that this mechanism is not an associative process which determines which memories are activated by specific cues in the environment, but rather it is a mechanism that filters out memories which do not belong to the ongoing reality (Schnider, 2003). Moreover, patients hold a strong conviction about what they perceive as the ongoing reality, just like control subjects who are not able to consciously alter their concept of ongoing reality. Schnider and colleagues suggest that this strong conviction can be explained by the fact that the ongoing reality checking is a relatively early process (around 200-300 ms after stimulus onset), compared to the recognition of the content of a specific memory which occurs about 400-480 ms after stimulus onset (Schnider, 2003; Wahlen et al., 2011). These results suggest that by the time the content of a specific memory is recognized, its representation has already been adapted according to whether or not it relates to ongoing reality. So this can explain the strong conviction associated with representations which had been associated with the ongoing reality and the failure to question these representations.

Another model establishes a parallel between delusions such as the Capgras syndrome and confabulations (Metcalf et al., 2007; Turner and Coltheart, 2010). They share many similarities. They both involve false interpretations, which are not correctly evaluated, leading to their

adoption, despite their bizarre content. The authors thus suggest a two-factor model to explain confabulations as well as delusions. The first factor for confabulations would be a failure in the processing of contextual and source information associated with memories, imaginings and ideas. Patients would thus be unable to dissociate ongoing reality, real memories and imaginings. This first factor is therefore different in patients producing confabulations and Capgras patients, and this explains the differences in the content of patients' subjective reports. However, this model suggests the existence of a second factor for explaining both delusions and confabulations: a failure to reject unsubstantiated claims. The existence of such an evaluation impairment in patients with confabulations has been highlighted by several theories (Burgess and Shallice, 1996; Gilboa and Moscovitch, 2002). This process of critical evaluation has been associated with the ventromedial PFC, which is often injured in patients with confabulations, but not in patients suffering from amnesia without confabulations (Gilboa and Moscovitch, 2002; Schnider, 2003). This second factor would explain why in delusions and confabulations, the false claims are experienced as true, associated with conviction and resistant to challenge. Turner and Coltheart (2010) propose that our belief evaluation system is organized in two sub-systems: an unconscious checking system and a conscious checking system. The unconscious checking system would process the information first and "tag" thoughts and interpretations that require additional conscious checking by the second sub-system. The presence of this "tag" would give rise to the feeling of doubt, triggering the conscious checking system and the absence of this "tag" would enable interpretations to directly affect behavior, without an additional checking step. A failure in this unconscious checking system would explain the strong conviction that patients have in their interpretations. Finally, according to this model, the conscious checking system involves reasoning, evaluation and monitoring processes, which can fail either because they are not correctly recruited, or because the strength of the conviction stemming from the failed unconscious checking system is such that conscious evaluation cannot be applied to delusional or confabulatory thoughts, or because they are impaired themselves.

Patients producing confabulations are therefore another very interesting example of our ability to create meaning and coherence in our experiences. So far, I have provided examples of patients suffering from neurological disorders who produce interpretations of their perceived reality and feelings, despite anomalous inputs. Note that this evidence is not anecdotal and is highly reproducible across patients. This series of examples highlights our ability to attribute meaning and create coherence in experiences related to pathological conditions. In the following section, I will provide striking examples of coherent and meaningful interpretations provided by control subjects in laboratory settings, suggesting that this ability is not specific to pathological conditions.

1.2. Evidence from control subjects

1.2.1. Attribution of animacy to geometric shapes

Interpretations are indeed not only found in patients; rather neuropsychology studies reveal extreme examples of these processes, which are also present in control subjects. One of the most illustrative examples of interpretations in control subjects has been provided by Heider and Simmel (1944). In this experiment, subjects were shown a moving-picture film with three very simple geometric figures, a big triangle, a small triangle and a circle. These figures were moving in various directions and at various speeds, around a rectangle, a part of which could move like a door (Fig. 1.3).

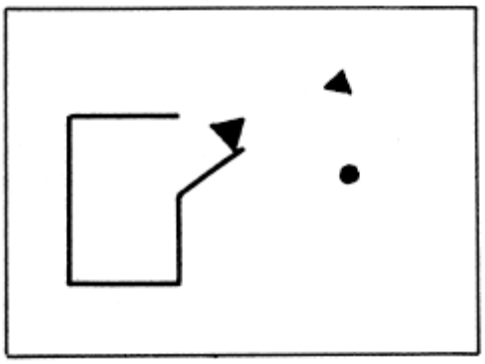

FIGURE 1.3: SCREENSHOT OF THE MOVIE PRESENTED BY HEIDER AND SIMMEL
Subjects were presented with simple geometric shapes, but reported complex interpretations of the movie, usually in the form of a connected story involving animate beings (adapted from Heider and Simmel, 1944).

Subjects were then asked to describe the film. Interestingly, all of the 34 participants except one interpreted the movements as actions of animate beings, mostly in a connected story. A typical example was: "A man has planned to meet a girl and the girl comes along with another man. The first man tells the second to go; the second tells the first and he shakes his head. Then the two men have a fight and the girl starts to go into the room to get out of the way and hesitates and finally goes in. She apparently does not want to be with the first man. The first man follows her into the room after having left the second in a rather weakened condition leaning on the wall outside the room. The girl gets worried and races from one corner to the other in the far part of the room. Man number one, after being rather silent for a while, makes several approaches at her; but she gets to the corner across the door, just as man number two is trying to open it. He evidently got banged around and is still weak from his efforts to open the door. The girl gets out of the room in a sudden dash just as man number two gets the door open. The two chase around the outside of the room together, followed by man number one. But they finally elude him and get away. The first man goes back and tries to open his door, but he is so blinded by rage and frustration that he cannot open it. So he butts it open and in a really mad dash around the room he breaks in first one wall and then another." Note that this elaborate story is the description of a movie involving two triangles and a circle, without any faces or body parts. Interestingly, when the film was presented in reverse order, subjects also interpreted the images presented, although the interpretations showed much more variation.

Similar movies have been presented to 12-month infants who already showed evidence of attribution of causal intentional states to simple geometric forms (Gergely et al., 1995). But this ability is affected in individuals with autism or Asperger syndrome: when describing movies involving simple geometric shapes in motion, their use of mental state descriptions was less extensive and less appropriate than in control subjects (Abell et al., 2000; Bowler and Thommen, 2000; Castelli et al., 2002)

This study is a perfect example of how we create sophisticated interpretations of our environment and attribute coherent meaning to the situations we observe, despite a very scarce input.

1.2.2. Influence of the context

A similar experiment showed the importance of context in the creation of interpretations (Wheatley et al., 2007). Subjects were presented with 12 animations of moving shapes, which were displayed against two different types of backgrounds: one biased the observer toward an animate interpretation and the other biased the observer toward an inanimate interpretation. Note that each subject only saw one background associated to a given moving shape and each moving shape was associated with a different background. For example, subjects were presented with a spinning shape either against a mountain background or against a playground background (Fig. 1.4). Then, they had to choose between four plausible interpretations of the shape: two animate (e.g. ice-skater, dancer) and two inanimate (e.g. crayon, spinning top). If their interpretation did not match any of the propositions, they could select the option "none of

FIGURE 1.4: INFLUENCE OF THE CONTEXT IN THE INTERPRETATIONS

Subjects were presented with the same moving shape overlaid on different backgrounds. On the left, the background is suggesting an animate interpretation (e.g. ice-skater), whereas on the right, the background is suggesting an inanimate interpretation (e.g. spinning top). The red dashed line is included to indicate the path of the shape, but was not actually visible (adapted from Wheatley et al, 2007).

the above". Interestingly, subjects did interpret the geometric shapes as animate or inanimate and these interpretations were successfully influenced by the experimental manipulation (88% agreement with the bias). Functional magnetic resonance imaging (fMRI) was used in this experiment and when the same geometric shapes were interpreted as animate rather than inanimate, activity in the "social network", including the lateral portion of the fusiform gyrus, the superior temporal sulcus, the medial prefrontal cortex, the posterior cingulate and the amygdala, was increased. This result suggests that activation in this circuit was not related to particular stimulus features, but rather to its conceptual representation.

This experiment shows that control subjects create interpretations that can be manipulated experimentally by changing the context in which they are presented.

1.2.3. Choice blindness

Another very striking illustration of the existence of interpretations in control subjects is termed choice blindness. This effect was found by Johansson et al. (2005). In this experiment, subjects were asked to choose between pictures on the basis of attractiveness. On some trials, immediately after their choice, they were presented with the picture they chose and asked to explain why they picked it. Unknown to participants, on some of these trials, the experimenter used a magic trick and presented the participants with the image they did not choose (Fig. 1.5).

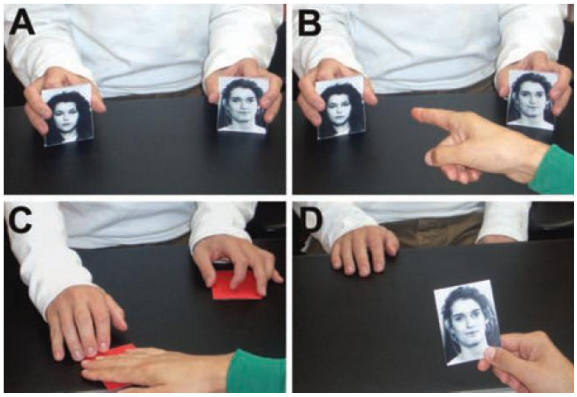

FIGURE 1.5: THE CHOICE-BLINDNESS EXPERIMENT

Subjects were presented with two pictures and asked to choose between them based on attractiveness (A). They indicated their choices by pointing (B). Unknown to participants, a second card depicting the opposite face was concealed behind the visible alternatives. On some trials, the experimenter flipped the pictures, so the feedback provided to participants was false (C). Subjects picked the picture and were asked to explain their choice (D).

Subjects were thus asked to justify a choice they did not make. Interestingly, no more than 26% of these manipulations were detected by the participants. In the other trials, subjects did not detect the change and thus offered an explanation for the choice they did not make. The reports in manipulated trials were delivered with the same confidence and the same level of detail as in non-manipulated trials. In manipulated trials, reports could be organized in different categories depending on the information they provided. Two categories were of particular interest: the "specific confabulation" category which contains reports referring to features unique to the picture the participants did not choose (e.g. "I chose her because she's radiant. I would rather

have approached her at a bar than the other one. I like earrings [whereas the woman they initially chose did not have earrings]!") and the "original choice" category which contains reports specifically referring to features only present in the originally chosen image, but not in the image they are currently watching (e.g. "Because she was smiling [said about the solemn one]").

This experiment highlights our capacity to build coherent interpretations: using a magic trick, it reveals a process that we probably normally use when asked to explain our behavior.

1.2.4. Dealing with uncertainty

Finally we are able to identify coherent and meaningful relationships among sets of random stimuli, as shown by Whitson and Galinsky (2008). In a series of 6 experiments, they provided evidence suggesting that people actively try to keep a sense of control over uncertainty by striving to find consistent relationships among stimuli, as a compensatory mechanism. Interestingly, in this series of experiments, lack of control was manipulated, in order to see how it interacted with illusory pattern perception. Lack of control was manipulated in two ways. In some experiments, subjects had to perform a concept-identification task: they were told that the computer would choose a concept and that they had to retrieve it. To do so, they were presented with pairs of symbols, in which one symbol correctly represented the selected concept and the other did not. Subjects had to decide which symbol was the correct one and were given feedback about their responses. In the lack of control condition, this feedback was random and independent from subjects' responses, so they were unable to correctly guess an answer. The second way of creating lack of control was to ask subjects to vividly recall an experience in which they lacked control over a situation. In one of the experiments, subjects were then presented with snowy pictures: half contained an embedded image and half were just random strokes. Participants in the lack-of-control group tended to see more meaningful patterns in the random strokes than participants in the control group. In another experiment, participants were asked whether they thought that an outcome and a potentially unrelated behavior (e.g. stomping his feet three times before entering a room and having his ideas completely ignored during the following meeting) were related. These links similar to superstitious beliefs were more frequent

in the lack-of-control group than in the control group. Finally, the relationship between lack of control and illusory pattern perception was tested in a financial domain (the stock market) by assessing if two uncorrelated sets of information are perceived as related. Sense of control was manipulated by describing the market as volatile or stable. Participants were presented with statements about two companies A and B containing either positive or negative information about their performance. The ratio of positive to negative statements was the same for both companies, but the amount of information seen for each company was different: company A had 16 positive and 8 negative statements, whereas company B had 8 positive and 4 negative statements. Participants were then asked to choose in which company they would invest and how many negative statements they remembered for each company. In typical illusory correlation paradigm, participants usually perceive a correlation between the infrequent behavior (here, negative statements which are less frequent than positive statements) and the group with less information (here, company B for which subjects received less statements) and thus tend to over-estimate the frequency of the two rare events occurring together. In the lack-of-control condition (facing a volatile market), participants invested less in company B and over-estimated the frequency of negative statements for this company, suggesting that they formed illusory correlation between the infrequent type of information (negative statements) and the infrequently presented group (company B).

This series of experiments suggests that experiencing lack of control leads to desire more structure in the environment. This can lead to the perception of illusory patterns as diverse as illusory figures in noise or spurious correlation in stock market information. Again this illustrates our propensity to seek patterns and consistent and meaningful information in random stimuli.

So far, the examples I have presented from both neurological patients and control subjects highlight our ability to create meaning and to seek consistency in our experiences. In the

following section, I will suggest a unifying framework which sets the stage for the work I present in this thesis.

1.3. Maximizing consistency: framework and objectives of the present work

All the studies presented so far share some similarities. Both patients and control subjects rely on their environment and their behavior to build coherent interpretations, despite the inconsistencies and uncertainty in the input they receive. The left hemisphere of split-brain patients creates meaningful explanations of their behavior, although it does not have access to the relevant causal information. Patients suffering from Capgras delusion or from confabulations generate coherent interpretations of their unusual inputs about face processing or reality monitoring, which fail to be identified as bizarre by their deficient belief evaluation system. This process of creating coherent interpretations based on the integration of information in time and space is also present in control subjects, as evidenced by different experiments. Interestingly, this propensity to create coherent interpretations is modulated by uncertainty.

Based on these different examples, it appears that interpretations, ranging from low-level visual perception to complex subjective reports, reflect the integration of sensory information with previous knowledge about the world. This relates to early accounts of perception as inference about sensory causes (Helmholtz, 1910). Indeed, what we define as meaningful interpretations are stories that are consistent with prior knowledge, given the information provided to the subject. For example, in the case of split-brain patients, they cannot report the information presented to their right hemisphere, but given the information available to their left hemisphere (the chicken claw presented on the right of the screen and the movement of their left hand toward the shovel), they provide an interpretation which integrates this information and previous knowledge about the world (how a shovel can be used for in the context of poultry farms).

A principled way to integrate sensory information with previous knowledge about the world is provided by Bayes' theorem. Accordingly, the probability of a cause A given a set of data B (the

posterior probability density function) is proportional to the probability of the set of data B given the cause A (the likelihood function) times the probability of the cause A (the prior probability density function). Interestingly, it has been observed that human observers behave as optimal Bayesian observers in numerous experiments (e.g Liu et al., 1995; Mamassian and Landy, 2001), integrating sensory inputs and prior knowledge about the world according to Bayes' theorem. This led to the 'Bayesian coding hypothesis', according to which the brain computes probability distributions and integrates priors and sensory information through Bayesian inferences (Knill and Pouget, 2004). The predictive coding framework has been suggested as a neurally plausible and computationally tractable way of performing these inferences (Rao and Ballard, 1999; Friston, 2005, 2010; Summerfield and Koechlin, 2008; Fletcher and Frith, 2009; Wacongne et al., 2011; Clark, 2013). According to this framework, brain areas are organized in a hierarchical manner and *conditional expectations* (the most likely cause of observed sensory information given that input, i.e. the expected value of the posterior probability density function) are generated at multiple levels of the hierarchy. These expectations are then projected backward to the immediately preceding level of the hierarchy and are used as empirical priors at this level. At the lower level, these priors are compared with conditional expectations and the prediction error is sent forward to the higher level (for an example see Fig. 1.6 in which the words "event" and "went" underlined in blue are represented by the exact same visual stimulus but can be dissociated using prior expectations about the meaning of the sentence).

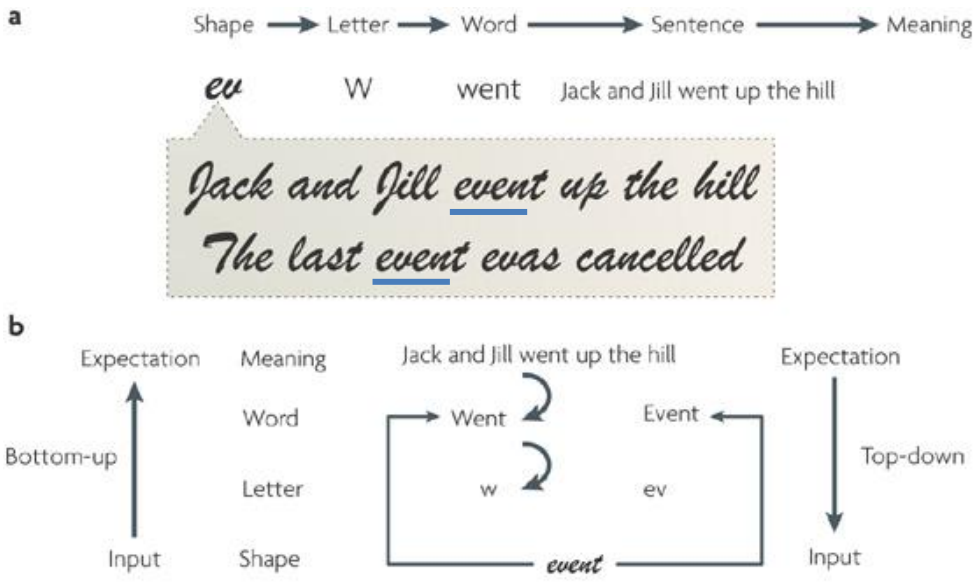

FIGURE 1.6: ILLUSTRATION OF A HIERARCHICAL BAYESIAN SCHEME FOR READING
To dissociate between the words "went" and "event" which are identical, a Bayesian hierarchical processing model is used. Here three level of the hierarchy are shown: at each level, the interpretation of the input to lower levels is constrained by expectations at higher levels (adapted from Fletcher and Frith, 2009).

I suggest that complex interpretations reported by patients or control subjects provide insight about inferences they made, by combining sensory data and their prior knowledge about the world. These complex interpretations can be viewed as the results of inferences in a high-level of the hierarchy in the Bayesian theories framework (Hirsh et al., 2013). For example, in the study of Heider and Simmel (1944), subjects were provided with a scarce input, so they attributed a higher weight to their prior knowledge and constructed the inference that the movie represents a connected story between animate agents. These inferences can be described as abductive inferences as defined by Charles Sanders Pierce, i.e. inference to the best explanation given the observed data. Interestingly, Coltheart and colleagues proposed that abductive inferences are at the basis of delusional beliefs such as those observed in Capgras patients and can be account for in a Bayesian framework (Coltheart et al., 2010).

One similarity shared by all the examples that I presented at the beginning of this thesis is that patients or control subjects provide meaningful and coherent interpretations of their

experiences. Consistency can be understood as a match between what is experienced and what is expected based on prior knowledge. The prediction is not always perfect and some inconsistencies can be encountered by subjects. An interesting question would be to understand how these inconsistencies are resolved when they arise. This resolution can be obtained through a modification of priors or by changing sensory samples (Friston, 2010). For example, the choice-blindness paradigm creates an inconsistency between the image that is presented to the subject and the image she is expecting to see based on her choice. When asked to explain a choice they did not make, subjects either referred to features specific to the person in the image they did choose (ignoring the sensory data coming from the new image) or referred to features specific to the new image (adapting their beliefs about their behavior).

Finally, a key element in the examples of coherent interpretations that I presented above is that they are the subjective reports of patients or control subjects about their experiences. These subjective reports are at the basis of the operational definition of conscious access used in cognitive science (Dennett, 1992; Dehaene and Naccache, 2001). Indeed, *being conscious of an information* is defined as being able to report this information. This transitive use of the word conscious should not be confused with its intransitive use which refers to the state of consciousness (e.g. "this patient is still conscious"), independently of the content of consciousness. Moreover, as I presented previously, interpretations involve integration of information across time and space, which has been suggested as a key feature of conscious processing in several theories of consciousness (Dehaene et al., 1998, 2006; Tononi and Edelman, 1998; Dehaene and Naccache, 2001; Tononi, 2004, 2008; Dehaene and Changeux, 2011). Therefore, in order to better understand how coherent interpretations are created, we should explore the link between this process and conscious access. Two questions arise. First, is the process of creating interpretations accessible to consciousness? Second, does the information used for building interpretations need to be reportable, or can interpretations be created using non-conscious information?

Taken together, all the examples presented in this introduction highlight our capacity to build coherent interpretations, even when we are confronted with inconsistencies. In my PhD work, I aimed at understanding how subjects update their priors in order to handle these inconsistencies and how this process interacts with conscious access. To do so, I combined experiments on novelty detection and cognitive control with experiments from social psychology, namely cognitive dissonance field. This original association of studies from different fields of cognitive neuroscience allows me to study how subjects update their knowledge about their environment, but also about themselves and their own preferences.

In the first part of my experimental work (Chapter 2 and Chapter 3), I analyzed the mechanisms of rule learning and how subjects identify and process stimuli contradicting these rules by updating their priors about the environment. To do so, I manipulated the statistical regularities of the stimuli presented to the subjects in order to create rules and explored the behavioral and electrophysiological responses to deviant stimuli, using scalp and intracranial electroencephalogram (EEG). I explored rules that can be identified automatically, as well as rules that require working memory over longer periods of time and the strategic use of subliminal and supraliminal stimuli. In the second part of my experimental work (Chapter 4 and Chapter 5), I investigated how subjects resolved inconsistencies in their own behavior by updating their interpretations to maintain internal consistency. To this aim, I used a classic experimental paradigm of the cognitive dissonance literature to create inconsistencies in subjects' behavior and analyzed how they change their preferences using behavioral and functional magnetic resonance imaging (fMRI) measures.

Therefore, in the following sections of the Introduction, I will first present a review of literature addressing the issue of rule extraction and novelty detection, but also cognitive control in response to conflicting information. Then, I will give an overview of the cognitive dissonance literature and how the different theories can inform the issue of interpretations.

2. How do we extract rules in our environment and process inconsistent stimuli?

In the first set of experimental studies presented in this thesis (Chapter 2 and Chapter 3), I was interested in how subjects update their priors based on regularities in the environment and how deviance from these regularities is processed. The link between prior updating and consciousness was also explored, by testing whether it is reportable and whether it can be based on non-conscious information.

To do so, regularities were established by manipulating the frequency of relevant events and responses to frequent and rare events were analyzed, using behavioral and EEG measures. Therefore, in this section, I will start by providing evidence that subjects are able to extract regularities in various cognitive domains. I will then review the literature on the signatures of novelty and conflict detection, focusing on evidence provided by EEG. Finally, I will examine previous studies on the interplay between these processes and consciousness.

2.1. Extraction of regularities in various cognitive domains

Our propensity to extract patterns or regularities in the environment has been demonstrated in very diverse cognitive domains. It relies on our ability to use transitional probabilities, i.e. to identify that for example, object A is always followed by object B but never by object C. Such extraction of regularities has been shown to improve subjects' motor performance. For example, Nissen and Bullemer (1987) designed a serial reaction time task in which in each trial a light appeared on a screen at one of four possible locations. Subjects had to press a button, out of four possible buttons corresponding to the four possible locations of the light. Interestingly, subjects were divided in two groups. The first group was exposed to a particular sequence of 10 trials, which was repeated throughout eight blocks of 100 trials. Conversely, in the second group of subjects, the location of the light was determined randomly for each trial. Note that subjects in the first group were not instructed about the existence of a sequence or regularity in the stimuli. They were only instructed to respond as fast as possible, without making errors. As shown in Fig. 1.7, the reaction times in this task were strongly reduced in the first group (continuous line)

compared to the second group (dotted line). In the first group, reaction times between the first and the eighth blocks decreased by 164 ms, while in the second group, only a decrease of 32 ms was observed. This result indicates that the learning of the motor sequence was strongly facilitated by the existence of a systematic regularity in the stimuli. Interestingly, when questioned, 11 out of 12 subjects indicated that they had explicitly noticed the sequence in the repeating condition (3 noticed it in the 1st block, 3 in the 2nd block, 3 in the 3rd block, one in the 4th block and one in the 5th block). They were reporting the sequence they identified by pointing to the different locations, rather than by describing the sequence verbally. This result reflects the motor feature of the learning involved in this task. Subjects were thus able to extract the relevant sequence of movements from the stimuli in order to improve their performance on this task. Even if they were not instructed to do so, they were actively looking for regularities in the trials to improve the accuracy of their predictions and to respond faster.

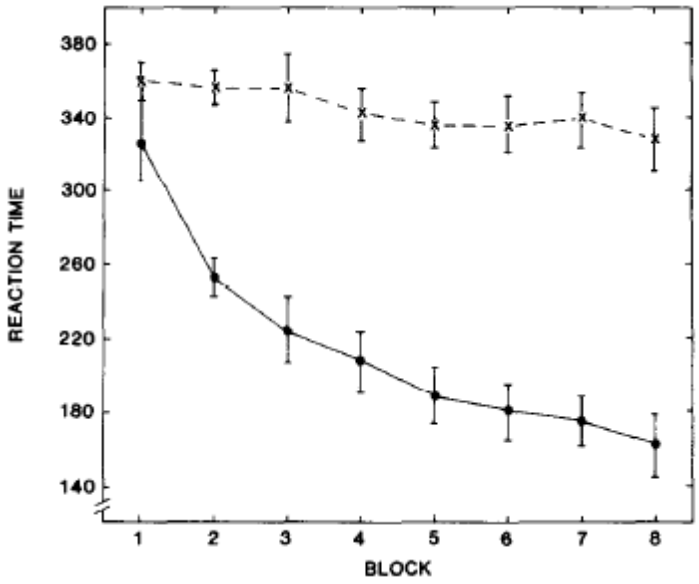

FIGURE 1.7: LEARNING OF SEQUENCES

Reaction times for subjects exposed to the repetition of a particular sequence of lights (continuous line) and subjects exposed to random sequences of light (dotted line). Adapted from Nissen and Bullemer (1987).

Evidence of regularity learning can also be found in visual tasks. Chun and Jiang (1998) identified a very interesting effect in visual search tasks: the contextual cueing effect. In this study, subjects performed visual search for targets (rotated T's) appearing among distractors

(heterogeneously rotated L's) that were displayed in invariant or variable spatial configurations and randomly intermixed within blocks. The "invariant" configurations were repeated across blocks and importantly targets always appeared at the same location in these configurations. Variable spatial configurations were not repeated across blocks. Each trial contained one of two possible targets and subjects had to press a response key corresponding to the presented target. The invariant configurations consisted of 12 randomly generated configurations which were repeated throughout the experiment, once per block. The location of the target was always the same in any given invariant configuration, but its identity was randomly determined to avoid priming of a particular response and to only assess the effect of the location predictability. The 12 variable configurations were randomly generated for each block. There were 30 blocks of 24 trials (12 invariant configurations, 12 variable configurations). Note that subjects were not instructed about the existence of invariant configurations. However, they were able to localize and discriminate targets more efficiently in invariant configurations than in variable configurations. This effect was found after only five blocks, i.e. after five presentations of the invariant configurations. Interestingly, subjects were not able to report that configurations were repeated. This benefit for target search performance in invariant configurations illustrates subjects' ability to identify regularities and patterns in the stimuli, although they did not consciously notice them.

Interestingly, this ability is developed very early by infants as evidenced by a study of Saffran et al. (1996). In this experiment, 8-month-old infants were familiarized with 2 minutes of a continuous speech stream consisting of four three-syllable non sense words (e.g. *bidaku*) repeated in random order. This stream was generated by a speech synthesizer, without any acoustic information about word boundaries. Therefore, the only available information to identify word boundaries was the transitional probabilities between syllable pairs, which were higher within words than between words. Then, learning was assessed by presenting each infant with repetitions of four three-syllable strings: two were "words" of the artificial language presented during the familiarization stage and two were "nonwords" which contained the

syllables heard during the familiarization but not in the order in which they appeared as words. During this test phase, infants controlled the duration of each trial, i.e. the duration of listening to each three-syllable string, by their sustained visual fixation on a blinking light. Interestingly, the infants showed a significant discrimination between "words" and "nonwords": they maintained fixation longer in trials during which "nonwords" were presented than in trials during which "words" were presented. This novelty preference revealed that 8-month-old infants are capable of extracting serial-order regularities in auditory streams after only 2 minutes of listening. In a second experiment, Saffran and colleagues (1996) refined this first result by showing that infants were able to discriminate between "words" and "part-words". Indeed, in this second experiment, "nonwords" in the test trials were replaced by "part-words", which consisted of the final syllable of a word and the first two syllables of another word. Thus these part-words did not correspond to words in the familiarization stream, but the discrimination between these part-words and words would reveal a deeper understanding of the boundaries between words and more generally of the structure of the auditory stream than the discrimination between nonwords and words. Infants showed a significant discrimination between the word and part-word stimuli, revealing their ability to extract information about the complex sequential statistics of syllables, even when provided with a very short familiarization.

These three examples provide evidence for our ability to extract regularities in our environment in a context in which this improves our performance and allows us to respond faster and more accurately. Interestingly, this propensity is so strong that we tend to extract regularities and patterns even when this is detrimental to our performance. For example, Yellott (1969) designed an experiment in which a light was flashed to the left or to the right on each trial. Before each trial, subjects had to predict where the light would appear. The location of the light was determined randomly and the most probable location was associated with a probability of 80%. Subjects' predictions matched the frequency of the actual presentation (frequency matching), i.e. subjects guessed the two lights with probabilities of 80% and 20%. Interestingly, in the last block, the light appeared wherever the subject predicted it would: if the subject

predicted that the light would appear on the left, the light would appear on the left, and vice versa. After 50 trials, the author stopped the experiment and asked subjects to comment on their impressions. Interestingly, several subjects reported that "they had solved the prediction problem near the end of the experiment" and they had found a pattern in the sequence of events. Of course these patterns were completely created by the subjects, but it highlights their propensity to seek patterns and regularities when interacting with an uncertain environment, especially when asked to provide predictions. This propensity leads to frequency matching, which is actually not an optimal strategy in this paradigm. Indeed, choosing the most probable option would lead to more correct guesses. For example, when the most probable option is presented in 80% of the trials, always choosing this option would lead to 80% correct answers, while frequency matching would lead to 68% correct answers (0.8*0.8+0.2*0.2). It has been shown that other animals tend to consistently choose the most probable option in such an experimental paradigm (Hinson and Staddon, 1983). These results indicate that human participants believe that there is a pattern in the sequence of stimuli, even when they are told that the sequence is randomly generated and this willingness to identify regularities leads to a non-optimal behavior. This experimental paradigm was used in split-brain patients in order to identify the role of the left and right hemispheres in this willingness to identify patterns (Wolford et al., 2000). In this experiment, patients were presented with sequence of stimuli in the left and the right visual hemifields. At the beginning of each trial, a row of three arrows (>>>) indicated for which visual field they had to make the prediction. Then they were asked to guess where a square would appear (either at the top or at the bottom of the screen). The top square was presented with a probability of 80% in the right visual field and with a probability of 70% in the left visual field. Interestingly, the patients showed two distinct behaviors for the two sequences of stimuli processed by their left and right hemispheres. When guessing about stimuli presented to the left hemisphere, patients matched the frequency of occurrence of previous stimuli. But when they guessed about stimuli presented to the right hemisphere, they used a strategy of maximization, i.e. they tended to always guess the most probable stimuli. This result

nicely highlights the link between the identification of patterns and regularities and the interpretations described in the first section of this Introduction, and suggests a crucial role of the left hemisphere in extracting patterns.

The examples presented above highlight our ability to extract regularities in our environment. This capacity is applied in various cognitive domains, e.g. motor learning, visual search and auditory speech perception. Interestingly, we tend to extract patterns even when the instructions clearly state that the sequences of events are randomly generated or when this strategy is not optimal. This suggests that the extraction of regularities in the environment is an automatic and implicit process. But in the studies by Yellott (1969) and Nissen and Bullemer (1987), participants reported explicitly the identification of a regularity in the environment. Thus, it is important to understand the interplay between the identification of regularities in the environment and reportability by addressing two questions. The first one is whether the extraction of regularities is always an automatic process. The second one is to what extent we need to be conscious of the regular stimuli to be able to extract the rule governing their presentation.

In the experimental studies conducted during my PhD, we addressed these questions using two paradigms in which subjects were presented with more or less frequent stimuli. This manipulation of stimuli frequencies allowed us to establish regularities in the environment. Note that for the sake of simplicity, the paradigms used in these studies did not involve complex patterns of regularity. In the first study (Chapter 2), I manipulated the frequency of sounds to create two levels of regularities in the auditory stimuli presented to the subjects (Bekinschtein et al., 2009): a "local" level of regularity within series of five sounds and a "global" level of regularity between series of five sounds. This first study allowed us to address the question of the automaticity of the extraction of regularities. In the second study (Chapter 3), I manipulated the frequency of trials including a conflict between the cue (a color word) and the target (a colored non-sense string) in a task derived from the Stroop task (Stroop, 1935). Trials including

a conflict are termed "incongruent trials" and trials in which the cue and the target point to the same response are termed "congruent trials". Classically, reaction times in incongruent trials are slower than in congruent trials, demonstrating a priming effect of the cue. In my experiment, subjects were presented with blocks consisting of 80% incongruent trials and I analyzed how this regularity in the environment influenced their responses, notably in incongruent trials. Moreover, the visibility of the cue was manipulated in order to test if the optimal strategy (i.e. preparing the response opposite to the cued one) could be applied even when subjects were not conscious of the conflict.

2.2. Neurophysiological signatures

Neural mechanisms related to the detection of regularities in the environment have classically been explored by the comparison of frequent and rare stimuli, also termed standard and deviant stimuli. In my experimental work, I also took this approach, therefore, in the following section, I will present a review of this literature, focusing on results in electroencephalography (EEG) as this is the method that I used in my experimental studies.

2.2.1. Novelty detection on short timescales: the mismatch negativity

Processing of regularities in the environment has typically been studied using the oddball paradigm (Squires et al., 1975). In this type of experiments, subjects are presented with repetitions of the same stimuli, randomly interrupted by the presentation of another rare stimulus. In most trials, the stimuli are thus highly predictable and it is interesting to compare the processing of these frequent and predictable stimuli ("standard stimuli") to the processing of rare stimuli, which violate the rule ("deviant stimuli"). This paradigm has been developed in the auditory domain, using changes in pitch or in tone to trigger rule violation, but has also been used in the visual, olfactory and somatosensory domains (Kekoni et al., 1997; Pause and Krauel, 2000; Pazo-Alvarez et al., 2003).

In this paradigm, the presentation of deviant stimuli typically results in an evoked potential that can be recorded by EEG or magnetoencephalography (MEG): the mismatch negativity (MMN), which has been described mostly in the auditory modality (Näätänen et al., 1978), although

analogous responses have been described in other modalities (Kekoni et al., 1997; Pause and Krauel, 2000; Pazo-Alvarez et al., 2003). The MMN is the component resulting from the subtraction of the response evoked by standard stimuli from the response evoked by deviant stimuli, it peaks around 100-200 ms after change onset and has a fronto-central distribution on scalp EEG (Näätänen et al., 2007). The MMN is elicited by any discriminable change in the sequence of sounds, e.g. changes in frequency, duration or intensity but also the presence of a silent gap instead of a tone. It thus reflects the detection of any violation of acoustic patterns or regularities (Näätänen et al., 2001). Source reconstruction studies have established that the MMN is generated by the superior temporal cortex bilaterally and possibly by the frontal cortex (Giard et al., 1990; Alho, 1995; Rinne et al., 2000; Opitz et al., 2002), as confirmed by fMRI studies (Molholm et al., 2005; Rinne et al., 2005). Intracranial recordings in human, which benefit from both high spatial and temporal resolutions, confirmed the localization of the MMN sources (Halgren et al., 1995a; Kropotov et al., 2000; Liasis et al., 2001; Rosburg et al., 2005). Note that the involvement of frontal regions has been related to involuntary attention switch following the detection of the deviance (Escera et al., 1998, 2003).

An interesting feature of the MMN is that it is thought to reflect an automatic novelty detection process (Näätänen et al., 2001). Indeed, it has been shown that the MMN can be detected even when the subject is not paying attention to the stimuli (Näätänen et al., 1978) or when the subject is actively engaged in another task (Bekinschtein et al., 2009). Moreover, it can be observed during rapid eye-movement sleep (Atienza et al., 1997), but also in patients in comatose state (Fischer et al., 1999; Naccache et al., 2005) or suffering from disorders of consciousness, i.e. in vegetative and minimally conscious states (Bekinschtein et al., 2009; Faugeras et al., 2011, 2012; King et al., 2013). This indicates that subjects do not need to be conscious of the stimuli to be able to detect violations in the regularity of the sound stream. Note that novelty detection in this case involves the comparison of the present stimulus to the previous stimulus to identify differences, so the detection of violation of regularities is related to changes in relatively short timescales. The automatic novelty detection indexed by the MMN is of

particular relevance in the context of this PhD, as I was interested in the issue of the reportability of regularity detection.

There are two competing hypotheses to explain the MMN. The first hypothesis is the model adjustment hypothesis (Winkler et al., 1996; Sussman and Winkler, 2001). According to this hypothesis, the MMN could reflect modifications of the perceptual model used for predicting auditory stimuli, based on mismatches between the predicted and actual auditory inputs. It has been proposed that temporal generators of the MMN underlie a sensory memory mechanism, while frontal generators would be involved in the comparison between the current stimuli and the memory trace (Giard et al., 1990). This hypothesis nicely fits into the more general predictive coding framework (Garrido et al., 2009; Wacongne et al., 2012). According to this model, perception arises from the integration of sensory information with priors about what can cause this sensory information. Prediction error is thought to be minimized at each level of the cortical hierarchy in order to estimate the most likely cause of the sensory information (Friston, 2005, 2010). In this framework, the MMN would reflect the existence of a prediction error when a deviant stimulus is presented, as this creates a mismatch between what is expected and what is experienced. Note that this relies on perceptual learning and on a change in the model of the world over repeated presentations of the standard stimulus to identify it as the most likely. The MMN is thought to reflect the existence of an update of subjects' priors about their environment and is the marker of the detection of a mismatch between their predictions and the inputs they receive. Wacongne et al (2012) developed an interesting biological model to address the issue of the implementation of predictive coding in the context of the MMN. The second hypothesis proposed for explaining the MMN is the adaptation hypothesis (Jääskeläinen et al., 2004; May and Tiitinen, 2010). This hypothesis is a simpler account of the MMN. It suggests that the MMN simply reflects a local neuronal adaptation or habituation, causing the delay and the attenuation of the N1 component observed in response to standard stimuli. This second hypothesis does not rely on a prediction mechanism. However, this model fails to account for some experimental observations, such as the existence of an MMN when a silent gap is presented instead of the

frequent sound (Yabe et al., 1997), but also the MMN response observed the same sound is repeated twice (AA) when subjects are used to an alternating sequence of sounds (e.g. ABABABAB… - Wacongne et al., 2012). Garrido and colleagues proposed to reconcile the model adjustment hypothesis and the adaptation hypothesis within the predictive coding framework by postulating two levels of prediction error suppression (within trials and between trials) which involve changes in synaptic sensitivity and efficacy during perceptual learning (Garrido et al., 2008, 2009).

In sum, the MMN is an event-related potential (ERP), which reflects the prediction error associated with rare stimuli in a very regular context. It appears to be based on the comparison between the current stimulus and the previous one, to identify the regularity violation. This ERP reveals the differences between the processing of frequent and predictable stimuli and the processing of rare stimuli, which cannot be easily predicted. This response to deviant stimuli is interesting in the context of this PhD as it highlights how the regularities in the environment can change priors and how inconsistencies are processed. Note that the MMN is the signature of novelty detection on short timescales.

2.2.2. Novelty detection on longer timescales: the P300 component

On longer timescales, novelty detection seems to involve more strategic processes, as indexed by the P300 component (Sutton et al., 1965). This response to auditory novelty is a large positive waveform, peaking about 300 ms after change onset. Two subcomponents can be dissociated: the P3a and the P3b. The P3a has a fronto-central maximum and peaks between 220 and 280 ms. It has been reported in response to infrequent distracter stimuli inserted randomly in sequences of standard and deviant target stimuli. This component can be observed even when subjects' attention is diverted from the auditory stimuli (Squires et al., 1975). It has been associated with the reorienting of involuntary attention, having been triggered by processes underlying the MMN (Escera et al., 2001; Friedman et al., 2001; Ranganath and Rainer, 2003). Conversely, the P3b has a more posterior distribution in scalp EEG and spans from 250 to 600 ms after change onset. It is related to the processing of rare target stimuli embedded in

sequences of standard stimuli. The neural generators of the P3a and the P3b are imprecisely defined. They include the hippocampus and frontal and parietal cortices (Picton, 1992). The P3b component is related to the sensory mismatch detection indexed by the MMN, but differs in an important aspect: it is highly dependent on attention. It is only observed when subjects are actively engaged in the task of detecting deviant stimuli (Bekinschtein et al., 2009) or when they pay attention to the stimuli (Wacongne et al., 2011). To explain this result, the context-updating theory proposed that the P3b component reflects the attentional processes governing an updating of the stimulus representation when a new stimulus is detected (Donchin and Coles, 1988). Interestingly, P300 can be observed in other experiments than oddball paradigms, when they involve the detection of a rare target. For example, it has been reported in the attentional blink paradigm. Subjects are presented with a series of visual stimuli including target stimuli which require a response. When the stimulus onset asynchrony (SOA) between two targets is too short (between 200 and 500 ms), the identification of the first target typically hinders the detection of the second target, although the second target is easily detected when no response is required on the first target. Using this paradigm, Sergent et al. (2005) have proposed the P3b component as a neural signature of conscious access.

Interestingly, the P300 component has been explained in the predictive coding framework (Wacongne et al., 2011; Chennu et al., 2013). It has been interpreted as a disconfirmation of predictions. In these studies, the authors take advantage of an experimental paradigm designed by Bekinschtein et al. (2009), which includes two levels of auditory regularities. The first level of regularity is based on series of five sounds: the first four sounds are always the same and the fifth sound can either be the same or be different. Violations of this first level of regularity elicit a MMN response. The second level of regularity is based on the global context in which series of five sounds are presented: the series of five sounds itself can either be frequent or rare. Rare series of five sounds trigger a P300 component. This experimental paradigm is interesting as it allows the dissociation between the MMN and the P300 and the testing of differences between the processes underlying these two components. In the predictive coding framework, the

difference between the MMN and the P300 is that they reflect different levels in the hierarchy of prediction errors. Indeed, the P300 is a signature of violations of regularities at the global level, relying on information stored in working memory, whereas the MMN reflects novelty detection on shorter timescales. This interpretation of the P300 in the predictive coding framework is in agreement with the context-updating theory (Donchin and Coles, 1988).

The dissociation between the MMN and the P300 is particularly relevant to the issue of the link between consciousness and regularity identification. Indeed, the MMN reflects an automatic process, whereas the P300 indexes a more strategic process, relying on working memory. In my experimental work (Chapter 2), I used the paradigm developed by Bekinschtein et al (2009) to better understand the neural mechanisms underlying novelty detection on local and global levels and to dissociate automatic and strategic processes.

2.2.3. Cognitive control

Finally, in the second experimental study (Chapter 3), I manipulated the frequency of response conflict in blocks of trials and assessed how this frequency could influence subjects' adaptation to conflict, which is a form of cognitive control. Cognitive control can be defined as the set of processes which allow flexible adjustment of our behavior to achieve our goals in the current environment. A classic example of experimental paradigm involving response conflict is the Stroop task (Stroop, 1935). In this task, subjects are presented with a color word (e.g. red) printed with a colored ink and they are asked to respond to the color of the ink, independently of the word meaning. There are two types of trials: "congruent trials" in which the word meaning (the irrelevant feature for the task) and the color of the ink (the relevant feature for the task) match (e.g. red printed in red) and "incongruent trials" in which there is a mismatch between the word meaning and the color of the ink (e.g. red printed in green). In incongruent trials, there is thus a response conflict, as the meaning of the word and the color of the ink point towards two different responses. Subjects are typically faster when responding in congruent trials than in incongruent trials. This effect on reaction times is called the Stroop effect. In this task, cognitive control processes are apparent in what is termed "conflict adaptation". Accordingly, the Stroop

effect is modulated by the context: it is smaller when incongruent trials are frequent in blocks of trials (Logan and Zbrodoff, 1979) or following an incongruent trial (Gratton et al., 1992). Conflict adaptation in blocks of trials reflects our ability to extract regularities in the environment to change our response strategies: when incongruent trials are frequent, the best strategy is to focus attention on the task-relevant feature and to ignore the task-irrelevant aspect in order to reduce conflict.

Conflict adaptation is therefore an interesting process to explore in the context of this PhD as it reflects a change in subjects' interpretations of their environment, based on the regularities they identified. Moreover, it can help address the issue of strategic use of supraliminal and subliminal information. I will now review here the literature on the neural signatures of conflict detection and conflict adaptation.

Activations in the anterior cingulate cortex (ACC) have been associated repeatedly with response conflict using positron emission tomography (PET) and fMRI (Pardo et al., 1990; Carter et al., 2000; Barch et al., 2001; Botvinick et al., 2001, 2004; Ridderinkhof et al., 2004b), for example in incongruent trials in the Stroop task (Banich et al., 1991; Bench et al., 1993). According to the conflict monitoring theory (Barch et al., 2001; Botvinick et al., 2001, 2004), the ACC is involved in the evaluation of the demand for cognitive control by monitoring response conflict. Note that the ACC is not directly involved in cognitive control, but rather it evaluates conflicts in information processing and triggers top-down control. This distinction has been highlighted by studies dissociating conflict and control. For example, Carter et al (2000) used an experimental design derived from the Stroop task. They used two types of blocks: mostly incongruent blocks (80% incongruent trials) and mostly congruent blocks (20% incongruent trials). These blocks differ in the degree of conflict expectancy. In mostly incongruent blocks, conflicts are very likely, so subjects need to exert high control on information processing and reduce the influence of the irrelevant feature of the stimuli (low conflict for incongruent trials). Conversely, in mostly congruent blocks, conflicts are unlikely, so subjects can rely more on the

irrelevant feature of the stimuli (high conflict for incongruent trials) and exert less control. These two conditions allow the comparison between high-control/low-conflict incongruent trials and low-control/high-conflict incongruent trials. Interestingly, this comparison showed that the ACC is more activated in low-control/high-conflict incongruent trials than in high-control/low-conflict incongruent trials, highlighting the role of the ACC in conflict evaluation, but not in strategic control. In EEG, the activation of the ACC has been associated with the N2 component. This component has a fronto-central distribution and peaks between 200 ms and 350 m after stimulus onset (van Veen and Carter, 2002; Nieuwenhuis et al., 2003; Folstein and Van Petten, 2008). This component is modulated by trial congruence (Danielmeier et al., 2009; Forster et al., 2010; Clayson and Larson, 2011). This EEG component is thus a marker of conflict detection.

Cognitive control processes triggered by this conflict detection have been related to the P300 component that I presented in the previous section. In experiments on conflict adaptation, it has been shown that this component is also modulated by congruence, although its functional significance is less clear than the N2 component. Indeed, it has been related to several processes, such as response inhibition (Frühholz et al., 2011) or response evaluation (Kopp et al., 1996).

2.3. Interplay between these processes and consciousness

An interesting issue is to understand how these different processes are related to consciousness. I have already mentioned a few points related to this issue that I will develop now. The first important question is what type of regularity can be identified automatically and what characterizes strategic rule learning. The second question is whether behavioral adaptation following the identification of regularities in the environment can be observed when stimuli remain unconscious.

2.3.1. Is the extraction of regularities related to consciousness?

The MMN and the P300 components are both observed in response to violations of auditory regularities but, as I presented before, they reflect two different processes. The MMN is associated with automatic novelty detection and can be observed when subject's attention is

diverted from the auditory stream (Bekinschtein et al., 2009 - Fig. 1.8 bottom left). This response to deviant stimuli is present in altered states of consciousness, such as sleep (Atienza et al., 1997), anesthesia (Koelsch et al., 2006) and in patients in comatose or suffering from disorders of consciousness (Fischer et al., 1999; Naccache et al., 2005; Bekinschtein et al., 2009; Faugeras et al., 2011, 2012; King et al., 2013). Conversely, the P300 component, and more precisely the P3b, is thought to reflect the deployment of selective attention (Fig. 1.8 top right) to task relevant stimuli and their subsequent conscious processing (Donchin and Coles, 1988; Kok, 2001; Sergent et al., 2005). This component has been used in clinical settings to assess patients' state of consciousness: the presence of a P300 was a reliable indication that the patient was conscious (Bekinschtein et al., 2009; Faugeras et al., 2011, 2012; King et al., 2013; Sitt et al., 2014).

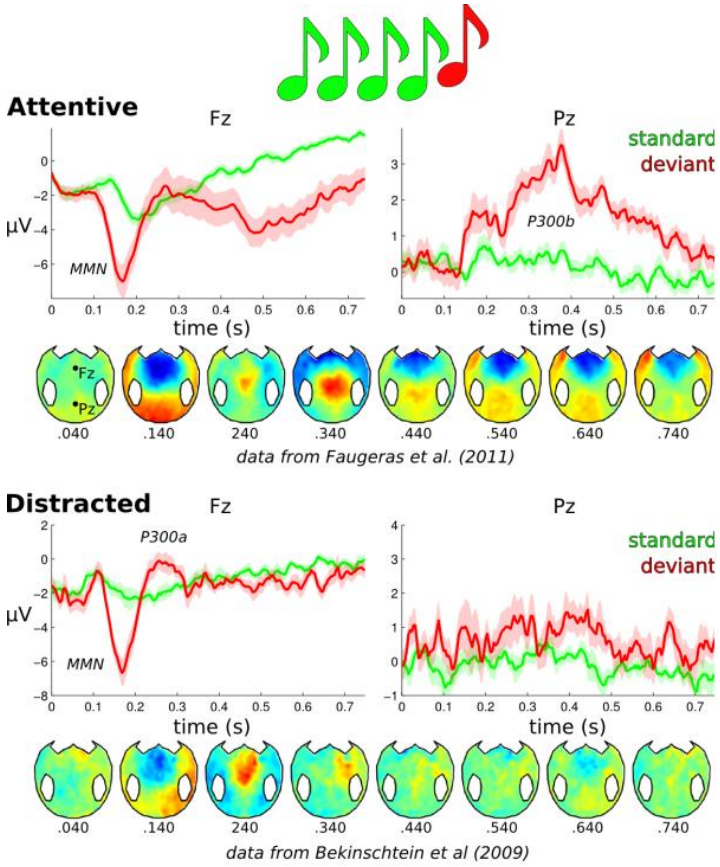

FIGURE 1.8: IMPACT OF ATTENTION ON THE MMN AND THE P300

In the oddball paradigm, deviant stimuli (red) elicit an MMN and a P300. Unlike the P300, the MMN is observed even when subjects do not pay attention to the stimuli (bottom). Adapted from King (2014).

An interesting feature of the MMN and the P300 is that they are not related to the same type of regularity. The MMN is associated with the detection of regularity violations on short timescales (from one sound to the next in a continuous auditory stream), which the P300 has been link to novelty detection on longer timescales (up to several minutes), which relies the comparison of current stimuli to information stored in working memory.

In the experimental work presented in this PhD (Chapter 2), we took advantage of the dissociation between these two components to explore the detailed mechanisms underlying novelty detection on short and long timescales, using intracranial EEG recordings. This first study aimed at understanding how subjects process stimuli which are inconsistent with rules established on a local or global context.

2.3.2. How does cognitive control relate to consciousness?

We were also interested in how subjects adapt their behavior after identifying regularities in their environment and how stimuli which are inconsistent with these regularities affect this new behavior. These issues are typically related to cognitive control functions and a highly debated topic is whether these functions can be applied on non-conscious stimuli.

Indeed, cognitive control has long been thought to be reserved for consciousness (Dehaene and Naccache, 2001; Jack and Shallice, 2001), but recent studies have challenged this view. According to Kunde et al. (2012), cognitive control experiments can be divided in two categories. The first category includes tasks triggering cognitive control functions by explicit events, e.g. task switching paradigms, Go/NoGo tasks or tasks involving cued orientation of attention. The second category includes tasks invoking cognitive control through implicit events, such as conflict frequency or conflict recency. In this type of tasks, there is no explicit cue indicating the need for cognitive control, but rather this need has to be identified based on the integration of information across trials.

Interestingly, in the first category of tasks, triggering cognitive control seems possible even when the explicit event is made non-conscious. For example, Lau and Passingham (2007)

designed a task switching experiment in which subjects were presented with words. When the instruction figure was a square, they had to decide whether the word was bisyllabic. When the instruction figure was a diamond, they had to decide whether the word referred to a concrete object. Before the instruction figure, a prime (either a square or a diamond) was presented, which was either clearly visible or masked. A congruency effect between the prime and the instruction figure was observed when the prime was visible, but also in the low visibility condition. This result shows that task switching, an instance of cognitive control, can be triggered by non-conscious stimuli. Note however that the instruction figure needs to be learned consciously before it can influence cognitive control functions non-consciously.

The evidence for cognitive control triggered by non-conscious stimuli is much more disputed in the second category of tasks. A first example of cognitive control triggered by an implicit event is conflict adaptation related to conflict recency, i.e. when cognitive control increases just after a conflict trial. This phenomenon has been studied by Kunde (2003). In this study, adaptation to conflict was only observed after conscious trials, suggesting that a non-conscious conflict cannot trigger cognitive control adaptation for the following trial. However, these results have been challenged by studies showing adaptation to conscious and non-conscious conflict (van Gaal et al., 2010; Desender et al., 2013). Conflict adaptation following non-conscious trials can be explained in two ways: either by the existence of non-conscious cognitive control processes, or by a different conflict experience in congruent and incongruent trials, related for example to slowing reaction times or to increase in error rates (Jáskowski et al., 2003; Kinoshita et al., 2008; Ansorge et al., 2011). The second hypothesis is supported by an experiment conducted by Desender et al. (2013) in which conflict experience was assessed in each trial. A second example of cognitive control triggered by an implicit event is conflict adaptation related to conflict frequency, i.e. cognitive control increases when conflict trials are frequent in order to reduce the number of errors. In a series of experiments, Merikle and colleagues provided evidence that subjects showed strategic responses based on conflict frequency only when the conflict was consciously perceived (Merikle and Cheesman, 1987; Merikle et al., 1995; Merikle and Joordens,

1997). This result was challenged by several studies showing blockwise adaptation to unconscious conflict frequency (Bodner and Masson, 2001; Jáskowski et al., 2003; Klapp, 2007; Bodner and Mulji, 2010). Note that these results could be related to conflict experience (Jáskowski et al., 2003).

Conflict adaptation provides an interesting framework to study how subjects adapt their behavior to regularities in the environment. Given the ambiguous results on the role of consciousness in conflict adaptation, we decided to study this issue by manipulating conflict frequency in a paradigm derived from the Stroop task. We included a novel condition allowing us to test the existence of strategy transfer: the strategic use of the conflict trials could be learned on conscious stimuli and then applied to non-conscious trials.

2.4. Summary

- *We constantly extract patterns in our environment.*
- *Regularity learning corresponds to an update in subjects' priors about the world.*
- *The processing of inconsistent information relative to extracted regularities provides insights on pattern identification mechanisms.*
- *The MMN and the P300 index novelty detection on different timescales.*
- *The MMN reflects an automatic process, whereas the P300 reflects a process involving working memory and selective attention.*
- *Behavioral adaptation after the identification of regularities in the environment is related to cognitive control processes.*
- *It is still unclear whether adaptation to frequent conflicts can be achieved when conflicts remain non-conscious.*
- *The N2 and the P300 are interesting EEG markers of conflict detection and subsequent cognitive control processes.*

3. How do we adapt our behavior to keep internal consistency?

In this PhD, I was not only interested in how subjects process inconsistencies in their environment, but also in how they deal with inconsistencies in their own behavior. According to the theory of "cognitive dissonance" proposed by Festinger (1957), subjects strive towards internal consistency and when they encounter inconsistencies, or to put it differently, if they hold two cognitions that are dissonant, they tend to reduce this dissonance. Therefore, this theory provides an interesting framework for understanding the way subjects adapt their behavior when facing lack of coherence in their attitudes. In the following sections, I will present the different experimental paradigms that have been used to study cognitive dissonance and I will focus on the free-choice paradigm that I used in my experimental work (Chapter 4 and Chapter 5). Then, I will review the different theories of cognitive dissonance, in order to identify the mechanisms proposed to be involved in cognitive dissonance resolution.

3.1. Different experimental paradigms to study cognitive dissonance

3.1.1. Belief disconfirmation

The cognitive dissonance theory was initially supported by the study of a group called the Seekers (Festinger et al., 1956). This group was led by Mrs. Keech who had received messages from aliens warning her about a big flood that would bring the end of the world before dawn on December 21, 1954. All the people would perish in this disaster except those who believed in the prophecy and who would be saved. The Seekers strongly believed in this prophecy and were very committed to it: as the December day approached, they quit their jobs, sold their possessions and even left their spouses who did not share their belief. The messages from the aliens were very specific: the group had to gather at Mrs. Keech's home on the evening of December 20 and wait for the spaceship that would take them away from danger.

Festinger and his colleagues predicted that on the morning of December 21, when the sun rose, as the spaceship failed to appear and the expected flood did not happen, the group of believers should experience an inconsistency between their beliefs about the end of the world and the

very existence of the world. Given their strong commitment to the prophecy and the elaborate preparation for this day, the Seekers were expected to be in the uncomfortable state of cognitive dissonance and to seek ways to reduce it.

To test this hypothesis and the theory of cognitive dissonance, one of the researchers, Stanley Schachter, infiltrated the group. He observed their preparation as well as the series of events happening right after midnight on December 20. The group gathered in Mrs. Keech's home and waited for the arrival of the spaceship. They had followed all the instructions from the aliens very precisely. When midnight arrived, there was no spaceship. But one of the group members realized that his clock still read 11:55, so they reset their watches. However, at 12:05, even on the newly set watches, there was still no spaceship. Then, one member suddenly realized that he did not fulfill one of the requirements: he had a metal filling in a tooth, although the aliens insisted that all metal objects should be removed. He removed it, but no spaceship arrived. The prophecy thus seemed to be disconfirmed. However, shortly past 4.00 am, Mrs. Keech received her final message: "This little group, sitting all night long, has spread so much goodness and light that the God of the Universe spared the Earth from destruction". This message provided the opportunity to restore consistency between the group's belief in doomsday and the absence of spaceship and flood. Interestingly, after they had generated the belief that their actions saved the world, they looked for social support for their story and wanted others to see that their actions had not been in vain and that the prophecy had not been disconfirmed.

In agreement with the cognitive dissonance theory, the Seekers were driven to find a way to restore their consistency, when facing a major discrepancy between their beliefs and observations of reality. They adapted their beliefs and behavior to make sense of what they had done and reduce their state of cognitive dissonance.

Following this initial observation, several experimental paradigms were developed to test the cognitive dissonance theory.

3.1.2. Effort justification

A first example is the effort justification paradigm designed by Aronson and Mills (1959). They reasoned that the suffering that goes into a given activity is inconsistent with people's desire not to suffer. According to the cognitive dissonance theory, enduring punishing activities should increase the positivity of our attitudes toward these activities, as this positive evaluation would decrease the dissonance between doing these activities and our desire to avoid pain. People who go through a lot of effort to obtain something would thus tend to value it more than people who reach the same goal with less effort. Note that this prediction is in contradiction with a common sense belief that punishment and suffering would produce negative reactions.

To test this prediction, Aronson and Mills (1959) asked female college students to participate in group discussions about the psychology of sex. Participants were randomly assigned to the Severe initiation group, the Mild initiation group or the Control group. When they arrived to the first meeting, they were told by an experimenter that in order to facilitate group discussions on this sensitive topic, participants were going to be seated in different rooms, communicating via headphones and a microphone. Participants in the Control group would then listen to the discussion directly. Conversely, in the Severe and Mild initiation groups, the experimenter explained that he would give the participant an "embarrassment test" in order to make sure that the subject could become a member of the group and discuss the topic without being too shy. In the Severe initiation group, subjects had to read twelve obscene words and two vivid descriptions of sexual activity in front of the experimenter. In the Mild initiation group, they had to read five words related to sex but that were not obscene, e.g. virgin. Following this test, participants were then asked to listen to a group discussion, but could not participate as they did not do the necessary reading prior to the meeting. This discussion was actually previously recorded, which means that all subjects listened to the same recording of "one of the most worthless and uninteresting discussions imaginable", as the authors describe it. Finally, participants were asked to rate the discussion and the participants. Interestingly, participants in

the Severe initiation group rated the discussion and the participants higher than participants in the Mild initiation and Control groups, although they had listened to the exact same recording.

This result is in agreement with the cognitive dissonance theory. Indeed, participants in the Severe initiation group hold dissonant cognitions: they had been embarrassed and the conversation was boring. One way to reduce the dissonance, or justify their embarrassment, is to increase the evaluation of the discussion.

3.1.3. Induced compliance

Another paradigm, called the "induced compliance paradigm", has been developed to test the cognitive dissonance theory, by Festinger and Carlsmith (1959). In this study, subjects were first asked to put twelve bobbins onto a tray, then to empty the tray and to refill it, and so on, with only one hand. They did this task for half an hour. Then, they were presented with a board containing 48 square pegs and they had to turn each peg a quarter turn clockwise and then another quarter turn and so on. They did this second task for another half hour. This first part of the experiment was designed to provide all subjects with an experience about which they would have a negative opinion. The experimenter then explained to the participants that they would have to convince another subject that the tasks they just did were actually interesting and fun. Depending on the condition, they were offered one or twenty dollars to do so. They spoke to the supposed other subject for two minutes, presenting the experiment as enjoyable. Then, they were taken to another room for an interview about their experience. They were asked whether they found the tasks interesting and enjoyable, whether the experiment gave them the opportunity to learn about their ability in these tasks, whether the experiment was measuring something important and whether they would want to participate in other similar experiments.

The hypotheses behind this design were as follow. After pretending that the experiment was enjoyable and fun, the subjects held two dissonant cognitions: they performed boring tasks and they pretended that it was interesting. However, this lie was consonant with the fact that they received a reward to perform it. The total magnitude of dissonance is related to both consonant

and dissonant cognitions and it decreases when the amount of consonant information increases. So, in this case, the total magnitude of dissonance decreases when the reward obtained for pretending that the experiment was interesting increases. Therefore, the group of subjects who received a small reward (one dollar) should tend to reduce their cognitive dissonance more than the group of subjects who received a bigger reward (twenty dollars).

Indeed, the final evaluation of the experiment suggested this pattern. Participants in the one-dollar group rate the experiment as more enjoyable, more important scientifically and were more willing to participate in similar experiments than participants in the twenty-dollar group. Interestingly, this pattern was not observed for the question about how much they learned which was not related to what they said to the supposed other subject.

During my PhD, we designed a within-subject induced-compliance experiment. Participants were asked to perform a boring task and they had to answer a series of 50 questions about the task itself and about the environment (e.g. how comfortable was the chair). Importantly, we interrupted their answers after 11 questions and explained to them that another subject for a different group arrived and that they had to pretend that the experiment was actually interesting, as this new subject could see their answers. We then restarted the series of 50 questions and before each question we presented the reward (either 1 euro or 1 cent) they would get for suggesting that the experiment was interesting. The reward depended on the category of the questions (e.g. 1 euro for each question about the task and 1 cent for each question about the environment) and this was counterbalanced across subjects. Finally, subjects were asked to faithfully answer all the questions on a different computer. The aim was to compare the attitudes of the participants before and after the manipulation, for the high- and low-reward categories. We obtained data from 17 subjects, but one was them did not believe the cover story, so we analyzed the data of 16 subjects. We computed a normalized cognitive dissonance index, by taking the difference in ratings between the first and the last evaluations divided by the amount of lie as measured by the difference in ratings between the first and the

second evaluations. This index was significantly different between high reward and low reward questions ($t(15) = 2.54$, $p = 0.02$). However, we were interested in studying the neural underpinnings of cognitive dissonance resolution, and with this experimental paradigm, the effect was fragile and only a few questions were analyzed in our within-subject design. Therefore, we decided to use another experimental setting: the free-choice paradigm, which was better suited for neuroimaging.

3.1.4. Free-choice paradigm: the original study

The free-choice paradigm was initially designed by Brehm (1956). I will present the original study and then focus on recent results obtained using this paradigm.

In this experiment, subjects were first asked to rate the desirability of eight items: au automatic coffee-maker, an electric sandwich grill, a silk-screen reproduction, an automatic toaster, a fluorescent desk lamp, a book of art reproductions, a stop watch and a portable radio. Then, participants were offered to choose between two objects and the chosen object was their payment for participating in the experiment. There were three groups of subjects depending on the choice they were offered. One of the two objects was always a highly-rated article. In the High dissonance condition, the other object was nearly as desirable as the first object. In the Low dissonance condition, the second object was poorly rated. Finally, subjects were asked to rate the eight objects again.

According to the cognitive dissonance theory, two hypotheses can be made. First, choosing between two alternatives should create dissonance, which could be reduced by making the chosen item more desirable and the rejected item less desirable. Second, the dissonance should be greater when the two alternatives are equally desirable.

The comparison between the first and the second ratings revealed that subjects tended to rate the item they chose higher and the item they rejected lower compared to their initial ratings. This effect was evaluated by the difference between the change in ratings (rating 2 minus rating 1) for the chosen item and the change in ratings for the rejected item. This measure, termed

"spreading of alternatives", was higher for the High dissonance condition than for the Low dissonance condition. These results suggest that choosing between two similarly desirable items induces a state of cognitive dissonance, which subjects tend to reduce by changing their evaluations of the chosen and rejected items.

These results have been replicated many times since 1956 (e.g. Steele et al., 1993; Heine and Dehman, 1997), not only in healthy adult subjects but also in amnesic patients (Lieberman et al., 2001), 4-year-old children and capuchin monkeys (Egan et al., 2007, 2010). Brain mechanisms underlying the reduction of dissonance in the free-choice paradigm have also been studied (Sharot et al., 2009; Izuma et al., 2010; Jarcho et al., 2011; Qin et al., 2011; Kitayama et al., 2013). However, all this literature has been challenged by the work of Chen and Risen (2010).

3.1.5. Free-choice paradigm: debated results

The main claim of Chen and Risen is that the original free-choice paradigm could produce a spreading of alternatives without any change in true preferences, because choices provide additional information about participants' preferences (Chen and Risen, 2010).

Let us take the example of a classical free-choice experiment to understand their argument. In this type of experiment, there are three steps: first, subjects rate the different items, second they choose between similarly rated items (difficult condition) or between clearly separated items (easy condition), and finally they rate all the items again. The critical measure is the change in ratings. As shown in Fig. 1.9 adapted from Sharot et al. (2009), in the difficult (critical) condition, chosen items are re-evaluated as more desirable in the second rating than in the first rating, whereas rejected items are re-evaluated as less desirable in the second rating, giving rise to a significant spreading of alternatives. On the contrary, in the easy (non-critical) condition, no such change in ratings is observed. These results are classically interpreted as reflecting a change in true preferences: in order to reduce their state of cognitive dissonance, subjects are thought to change their preferences to rationalize their choices.

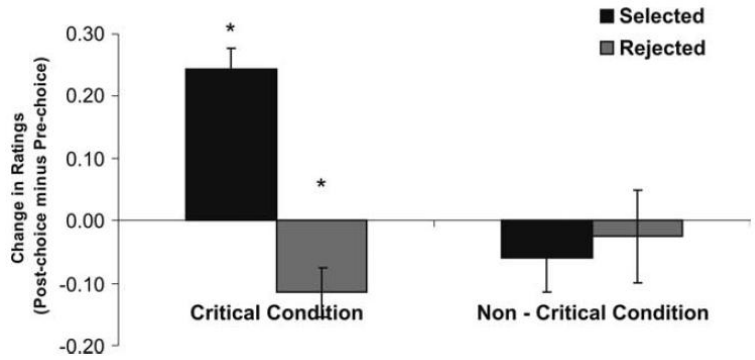

FIGURE 1.9: TYPICAL RESULTS OF A FREE-CHOICE EXPERIMENT

Following difficult choices (critical condition), selected items (black) are rated higher than before the choice, whereas rejected items (grey) are lower. This effect is not observed following easy choices (non-critical condition). Adapted from Sharot et al, 2009.

Interestingly, Chen and Risen showed that these results can partially be explained without any change in true preferences. They rely on three commonly accepted hypotheses. First, people's ratings are at least partially guided by their preferences. Second, people's choices are at least partially guided by their preferences. Note that this does not imply that subjects will always prefer their chosen item, but rather it implies that subjects are more likely to prefer the chosen item than the rejected item (i.e. choices are not random). Finally, ratings are not a perfect measure of preferences.

This last hypothesis implies that for each item, there is a distribution of possible reported preferences centered on its true preference (Fig. 1.10). Subjects are usually asked to provide discrete ratings, although the distributions of possible reported preferences are continuous. We can assume that ratings correspond to a sample of possible reported preferences rounded to the nearest integer. As shown on Fig. 1.10, for a pair of items A and B, the distribution of possible reported preferences can be partially overlapping. During the first rating (R1), items A and B can thus be rated equally: in the example presented in Fig. 1.10, R1(A) = R1(B) = 7. However, during the choice, according to the second hypothesis of Chen and Risen, the probability of choosing B is higher than the probability of choosing A, because the true preference for B is higher than the true preference for A. Finally, due to a regression to the mean effect, the second rating for A is

likely to be lower than the first rating, whereas the second rating for B is likely to be higher than the first rating: in the example presented in Fig. 1.10, R2(A) = 6 and R2(B) = 8. This statistical artefact is in the same direction as the effect of interest: spurious spreading of alternatives can be observed even in the absence of change in true preferences. Indeed, note that in this example, the distribution of possible reported preferences for A and B are exactly the same for the ratings 1 and 2, i.e. true preferences are the same in both ratings. Interestingly, in the easy choice condition, in which paired items are given very different ratings during the first rating, the choice does not provide any additional information. In this case, no spreading of alternatives is expected.

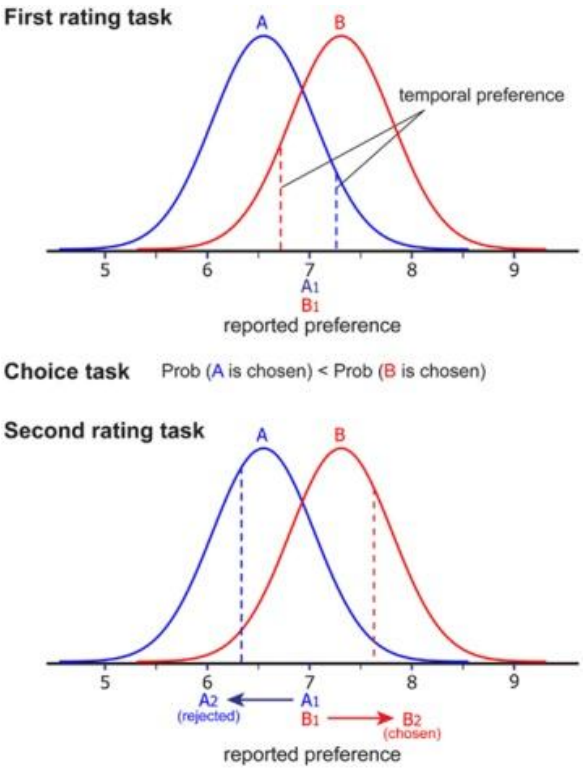

FIGURE 1.10: SCHEMATIC ILLUSTRATION OF CHEN AND RISEN'S CRITICISM
Each curve represents the distribution of subject's preference for item A (blue) and B (red). The same rating is given for both items during Rating 1. But given the true preferences for A and B, B is more likely to be chosen than A. Due to a regression to the mean effect, the second rating for A is likely to be lower than the first rating, whereas the second rating for B is likely to be higher than the first rating. This spreading of alternatives can be observed without any change in true preferences (adapted from Izuma and Murayama, 2013).

We ran simulations of free-choice experiments in order to control true preferences and to test the existence of spreading of alternatives without changing these true preferences. The advantage of simulated data is that real preferences are known. We assumed that for a given item, ratings were normally distributed around the real preference and that the standard deviation was the same for all items and for both ratings (before and after the choice). We coupled items with a similar procedure as in the free-choice paradigm: pairs of items were based on proximity in ratings during the first rating step. Finally, choices were modeled using scores normally distributed around real preferences with a standard deviation which was the same for all items: in a given pair, the chosen item was the one with the higher score. Real preferences were randomly chosen. The only parameters that were used in these simulations were the standard deviation for the ratings and the choice. These simulations suggest that chosen items tend to be under-estimated during the first rating, whereas rejected items tend to be over-estimated during the first rating, as choices reveal information about the underlying true preferences (Fig. 1.11 left). This clear separation tends to be cancelled in the second rating, as a consequence of the regression to the mean (Fig. 1.11 right).

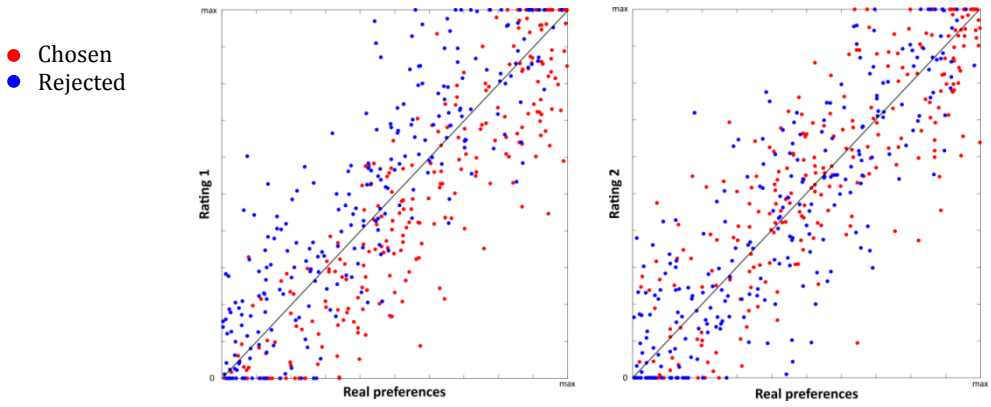

FIGURE 1.11: SIMULATIONS OF THE FREE-CHOICE EXPERIMENT

When classifying items as chosen or rejected in the first rating (left), it appears clearly that chosen items tend to be underestimated, while rejected items tend to be overestimated in this first rating. During the second rating (right), the segregation tend to disappear, highlighting the regression to the mean effect

Similar simulations have been performed by Izuma and Murayama (2013), which confirm the existence of spreading of alternatives without change in true preferences, only in the difficult choice condition, using 10 000 simulations (Fig. 1.12).

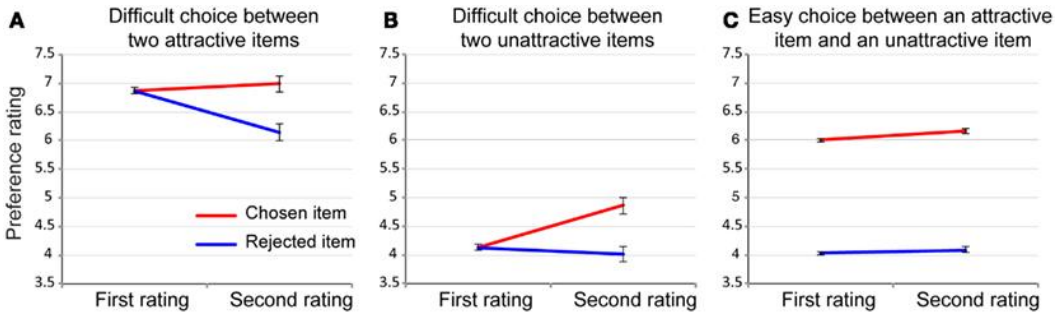

FIGURE 1.12: SIMULATED PATTERNS OF PREFERENCE CHANGE ACROSS DIFFERENT CHOICE CONDITIONS

Simulated reported preference changes following difficult choices between two attractive (left) and unattractive (middle) items and following easy choices (right)

This devastating argument challenged the results of previous studies, especially because it highlighted the fact that the control condition used so far (the easy choice condition) was not appropriate to control for the statistical artefact. Therefore, in their original paper, Chen and Risen propose a new control condition: the rate-rate-choose condition (Chen and Risen, 2010). In this condition, subjects are asked to rate items twice in a row and then to choose between pairs of items formed using the results of the first rating. In this condition, changes in ratings could only be attributed to the noise in ratings, and not to the choice, as it occurs after the second rating. This new control condition allows the estimation of the magnitude of the statistical artefact and a proper estimation of the effect of interest, i.e. choice-related preference change. Chen and Risen provide the results of two experiments using this control condition, which suggest that although much smaller, choice-induced preference change can be found when properly controlling for the artefact.

Following the identification of this artefact, several authors tried to address the problem (Egan et al., 2010; Izuma et al., 2010; Alós-Ferrer and Shi, 2012; Alós-Ferrer et al., 2012; Sharot et al., 2012; Johansson et al., 2013; Nakamura and Kawabata, 2013). First, Alós-Ferrer and Shi (2012)

demonstrated that the mathematical proof developed by Chen and Risen (2010) was not valid. Indeed, Chen and Risen tried to demonstrate that positive spreading of alternatives without change in actual preferences could be obtained in all the situations in which their three hypotheses were valid. But they could only show that at least one situation with positive spreading of alternatives without change in actual preferences exists. Importantly, this is sufficient to cast doubt on the experimental paradigm. Alós-Ferrer and Shi highlighted this error in the mathematical proof, but confirmed the importance of improved experimental designs, as the existence of positive spreading without change in actual preferences could be demonstrated in certain conditions.

Several improved designs have consequently been proposed. Some authors used the "rate-rate-choose" paradigm suggested by Chen and Risen (Chen and Risen, 2010; Izuma et al., 2010; Sharot et al., 2012). Other authors used a blind choice paradigm (Egan et al., 2010; Sharot et al., 2010; Nakamura and Kawabata, 2013): indeed, the problem highlighted by Chen and Risen is due to the fact that choices are not independent from preferences and reveal information about them. The idea is that if subjects make choices without knowing objects' identities, no additional information about subjects' preferences will be provided by these choices. Accordingly, simulation results showed that when the choice does not reflect participants' preferences, the paradigm does not produce artificial spreading of alternatives (Izuma and Murayama, 2013). This experimental design is therefore very interesting. However, the studies using it suffer from other methodological problems. In Egan et al (2010), infants and capucin monkeys are tested. In infants, attitude change is assessed by a second blind choice, which by definition do not provide information about preferences. In monkeys, preferences after the blind choice are measured using a series of 10 (non-blind) choices. But according to the cognitive dissonance theory, each of these choices should affect preferences, making the results of this experiment difficult to interpret. Sharot et al (2010) also used a very astute design: participants were told that the vacation destination names presented during the choice would be presented only very briefly and would be masked, so they would not be able to consciously perceive the names. In reality,

only nonsense strings were presented to avoid a potential effect of subliminal perception. Participants were asked to choose anyway and the pair of destinations associated to a given trial was then presented on the screen with a star indicating which destination had been "blindly chosen". Thus these blind choices could not reflect subjects' preferences and it is an elegant way to avoid the artefact highlighted by Chen and Risen. However, the results of this study are weak: the classical spreading of alternatives was not significant and only preferences for chosen items were significantly different in the second rating compared to the first rating (Sharot et al., 2010). Nevertheless, these results have been replicated by Nakamura and Kawabata (2013), who adapted the blind choice paradigm to facial images that subjects had to judge in terms of attractiveness. Using a similar approach, Johansson and colleagues used the choice blindness effect presented in the first section of this introduction to dissociate preferences from choices (Johansson et al., 2013). They found that participants tended to prefer the face that they believed they had chosen (while they actually rejected it), in agreement with the prediction of the cognitive dissonance theory. Interestingly, note that the change in preferences was thus related to the choice participants believed they did based on the feedback, but not on their actual choice. Finally, another strategy was proposed by Alós-Ferrer et al. (2012): the implicit-choice paradigm. Participants were asked to rate 80 vacation destination names, then to make choices between 16 item pairs and finally to rate again all items. But, contrary to the classical free-choice paradigm, the choice task included pairs of similarly rated items ("direct choices"), but also "implicit choices" which consisted of two choices including two similarly rated items *a* and *b* which were paired respectively with an item rated higher *h* and an item rated lower *l.* The idea of this paradigm is that assigning items *a* and *b* to pairs in which they have a high probability of being rejected and chosen respectively will allow their categorization without direct comparison, so without revealing any information about the actual preferences. This paradigm is interesting, but suffers from methodological problems as well. First, for most of their analyses, the authors selected only implicit choices in which subjects' behavior corresponded to their expectations (choosing *b* and rejecting *a*); this selection bias was corrected in the robustness

check test. Second, as noted by Izuma and Murayama (2013), this paradigm might not be completely immune to the Chen and Risen's artefact, as implicit choices could reveal information about preferences for *a* and *b* if the distance between *a*, *b, l* and *h* is too small. This argument seems to be validated by the effect of distance found by Alós-Ferrer et al. (2012).

The studies using experimental designs derived from the free-choice paradigm to correct for the statistical artefact pointed out by Chen and Risen (2010) are thus interesting but often suffer from other methodological issues. An interesting feature of these studies highlighted by Izuma and Murayama (2013) is that the effect size of choice-induced preference change when correcting for the statistical artefact is much lower than initially thought. It is thus important to re-establish some of the previous results and one particularly interesting issue is to understand the mechanisms of genuine choice-induced preference changes.

3.2. The theories of cognitive dissonance

Several theories have been proposed to explain cognitive dissonance effects and they can be classified into two categories according to the mechanisms thought to be involved in these effects. The first category includes Festinger's original theory and many of its revisions, which implicitly suggest that cognitive dissonance effects rely on high-level cognitive processes, such as memory and executive control. The second category includes mostly experimental studies showing that memory for example is not necessary for cognitive dissonance effects and thus suggesting that these effects are based on lower-level processes than initially thought. This distinction is particularly relevant for this PhD as these two categories of theories do not have the same implications on one of the issues I am interested in, namely whether subjects need to be conscious of the inconsistency to be able to adapt their interpretations. Therefore, I will organize my review of the different theories of cognitive dissonance in two parts.

3.2.1. Theories suggesting the involvement of high-level cognitive processes

The results obtained with these different experimental paradigms can be explained by the theory of cognitive dissonance, initially proposed by Festinger (1957). As mentioned previously, this theory postulates that when individuals hold inconsistent or dissonant cognitions, they tend

to reduce the uncomfortable state arising from this situation by changing their cognitions to reestablish consistency or consonance. The underlying hypothesis is that subjects try to establish consonance between their opinions, attitudes, knowledge and values, i.e. there is a motivational drive towards consonance among cognitions. Dissonance between two cognitions is defined by Festinger as the situation in which, when considering the two cognitions alone, the opposite of one of them follows from the other. It can be elicited in several situations. For example, when subjects make decisions, the positive characteristics of the rejected element and the negative features of the chosen elements are dissonant with the knowledge of the choice. Dissonance is also present when subjects are asked to publicly state an opinion that is contradictory to their own private opinion or when subjects receive new information that can be inconsistent with their previous cognitions. Interestingly, Festinger proposes that the magnitude of dissonance is related to the importance of the cognitions for the subject. The central claim of the cognitive dissonance theory is that the presence of dissonance triggers a motivational drive to reduce it and the strength of this drive is proportional to the magnitude of the experienced dissonance. Festinger suggests three ways to reduce this dissonance: changing one's cognitions (e.g. attitude changes in the free-choice paradigm), adding new consonant cognitions (e.g. when the Seekers looked for social support) or decreasing the importance of the dissonant elements (e.g. trivialization – Simon et al., 1995). Dissonance reduction is thought to be achieved through the modification of the least resistant cognition: it is easier to change a personal attitude than a publicly committed behavior. An interesting feature of this theory is its generality: dissonance can be defined for any type of cognitions, including attitudes, beliefs, perceptions and behaviors. Moreover, the theory of cognitive dissonance makes non-intuitive predictions, based on the magnitude of the experienced dissonance.

Although this is not clearly stated by Festinger, an underlying hypothesis in this theory is that explicit memory is involved in cognitive dissonance resolution, as the discomfort and the arousal need to be attributed to the discrepancy between the initial attitude and the counterattitudinal behavior (Lieberman et al., 2001; Jarcho et al., 2011). Festinger's formulation of cognitive

dissonance suggests that relatively slow and self-reflective processes are involved in the dissonance reduction. Note however that subjects are probably not aware of the rationalization process itself (Nisbett and Wilson, 1977), but this theory implicitly proposes that they need to be aware of the inconsistency in their behavior.

Following this original formulation by Festinger, several revisions have been proposed to account for the data provided by the different experimental paradigms presented above.

The original theory of cognitive dissonance was first challenged by Bem (1972). He indeed suggested that the experimental data supporting the cognitive dissonance theory could be more parsimoniously explained. This theory relies on two assumptions: first, subjects do not always have direct access to their own attitudes, beliefs or other cognitions, but instead they have to partially infer them from observations of their own overt behavior; second, as internal cues can be weak or ambiguous, subjects are in the same position as an observer who would rely on external cues (one's behavior and the context of this behavior) to infer one's internal states. Based on these hypotheses, Bem suggested that the experimental results obtained in the studies presented above could be explained by the fact that subjects are inferring their attitudes from their behavior. This explanation does not require the existence of a negative arousal state associated with dissonance that participants would try to reduce, as suggested by Festinger (1957). Instead, participants would simply infer their most probable attitude towards the task based on their overt behavior. For example, in the induced compliance study by Festinger and Carlsmith (1959), students were asked to tell another student that the task they just performed was fun and interesting, while it was actually boring, and received a reward of 1$ or 20$ for doing so. According to the self-perception theory, the results can be understood by taking the point of view of an external observer who would hear the positive statements of the participant about the task and who would know that he received either 1$ or 20$. This external observer would then be asked about the actual attitude of the participant: in the case of 1$ reward, he would exclude financial incentive as a motivating factor for stating that the task was fun and he

would infer that the participant enjoyed the task, while in the case of 20$ reward, he would not be able to infer the participant's personal attitude, as the reward would seem sufficient to explain his behavior independently of his personal attitude, and he would thus infer that the attitude of the participant is similar to the attitude of control subjects. Self-perception theory proposes that subjects themselves behave like this external observer and that they inferred their personal attitudes depending on their behavior and its context. This interpretation of the data does not require an aversive motivational state to change participants' attitudes, but only inferences based on one's behavior and potential controlling variables. Note however that these inferences must rely on explicit memory of the recent behavior (Lieberman et al., 2001).

Another revision of the cognitive dissonance theory suggests that conditions triggering a high dissonance arousal create a threat to participants' sense of self. Indeed, the theory of self-affirmation (Steele, 1988; Steele et al., 1993; Sherman and Cohen, 2006) suggests that we are motivated to see ourselves as good and honest people and that any evidence of the contrary will tend to make us rationalize our behavior in order to preserve the ideas we have about ourselves. So according to this theory, cognitive dissonance resolution is not triggered by any inconsistency between two cognitions, but rather only when inconsistencies threaten self-integrity. According to the self-affirmation theory, preserving self-systems is the primary influence that motivates the attempts to reduce cognitive dissonance. Indeed, in the different examples of cognitive dissonance experiments presented above, subjects are asked to lie to another student or to endure embarrassment in order to join a worthless group, and according to Steele and colleagues, these situations create a threat to subjects' self-systems. In order to reduce this threat and to preserve self-integrity, the self-affirmation theory suggests that subjects would affirm the global integrity of their self. Note that according to this theory, any behavior restoring their self-integrity would allow participants to reduce the threat to their self-system, not only the change in cognitions specifically related to the threat. The self-affirmation theory thus differs from the original cognitive dissonance theory, according to which only changes in cognitions specifically related to the inconsistency would allow a reduction of cognitive dissonance.

Another theory involving the self has been proposed by Aronson (1968): the self-consistency theory (Aronson, 1968; Thibodeau and Aronson, 1992), which emphasizes the need for the self to be involved in the inconsistencies triggering cognitive dissonance. Similarly to the self-affirmation theory, Aronson suggests that people need to see themselves as good and honest people, and that if cognitive dissonance exists, it is because the participant's behavior is inconsistent with his self-concept. In other words, the self creates expectations of how one should behave and if one's behavior is inconsistent with these expectations, it creates the state of cognitive dissonance. Note that the resolution of this negative arousal involves changes in cognitions specifically related to the inconsistency, contrary to what is assumed by the self-affirmation theory.

These two theories involving the self-concept in cognitive dissonance implicitly suggest that dissonance reduction is based on high level processes, such as explicit memory, as the inconsistency between one's attitude and one's behavior need to be identified. These theories did not differ from Festinger's proposal in terms of postulated mechanisms for dissonance reduction.

Cooper and colleagues developed the New Look model of cognitive dissonance (Cooper and Fazio, 1984; Cooper, 2007) to provide another explanation of the results obtained with the experimental paradigms presented above. This model separates two aspects of cognitive dissonance theory: first, the dissonance arousal generation and second the motivation for people to change their attitudes. One important difference between this theory and Festinger's cognitive dissonance theory is that in this case, the trigger of dissonance arousal is not inconsistencies per se, but the negative consequences of the behavior. Moreover, these aversive consequences should be irrevocable to induce cognitive dissonance and associated with a strong feeling of responsibility. According to this model, only the combination of irrevocable aversive consequences of one's behavior and a strong responsibility feeling for these consequences would lead to dissonance arousal. In the New Look view, attitude change following dissonance

arousal does not occur to restore consistency, but rather to render the consequences of behavior non-aversive. In order to trigger attitude change, dissonance arousal should first be labelled as negative and attributed to one's own behavior (as opposed to an external source) to give rise to a motivation to reduce it.

The New Look model of cognitive dissonance was further refined in line with the self-related theories. According the Self-Standards Model of Cognitive Dissonance (Stone and Cooper, 2001), a given behavior can be evaluated according to different standards in order to assess the averseness of its consequences. People can use normative standards to judge the outcomes of their behavior, i.e. they can assign to their own behavior the value that most people would assign. For example, most people agree that lying to another student is a bad thing and participants in Festinger and Carlsmith (1959)'s experiment can judge their behavior according to this normative standard of judgment. People can also use personal standards to judge the outcomes of their behavior, by referring solely to their own values, judgments and desires. These personal standards may or may not be similar to normative standards. The difference in the standards used for judging one's behavior explains in which conditions cognitive dissonance magnitude is related to self-esteem. Indeed, when behavior is judged according to personal standards, self-esteem should modulate the magnitude of cognitive dissonance, whereas when normative standards are used, no modulation of cognitive dissonance magnitude by self-esteem should be observed.

These two theories proposed by Cooper and colleagues heavily rely on reflective processes and although it is not explicitly mentioned, this framework also assumes that cognitive dissonance reduction involves high-level cognitive processes.

Finally, the meaning maintenance model (Heine et al., 2006; Proulx and Inzlicht, 2012; Proulx et al., 2012) interprets cognitive dissonance effects in a larger framework in which all inconsistency compensations studied in social psychology are understood as responses to violated expectations. This proposal is close to the initial formulation of cognitive dissonance by

Festinger (1957), but includes different compensation mechanisms, such as affirmation, accommodation, assimilation and abstraction, to understand how subjects handle the inconsistencies they encounter.

3.2.2. Cognitive dissonance resolution without explicit memory?

In the different theories presented above, the precise mechanisms of cognitive dissonance reduction are unclear. They provide different accounts of cognitive dissonance at the conceptual level, but they do not clearly state what cognitive processes should be involved in cognitive dissonance resolution, although their formulation suggests the involvement of reflective processes, notably explicit memory. A series of experimental studies of cognitive dissonance have challenged the view that high-level processes such as explicit memory or executive control processes are involved in cognitive dissonance resolution. Note however that no theoretical account has been developed to integrate these findings in a framework stating clearly the type of cognitive mechanisms being involved.

A first example is provided by Lieberman et al. (2001). In this study, the authors examine cognitive dissonance reduction in amnesic patients using the free-choice paradigm. In the first experiment, participants (12 patients and 12 control subjects) examined two sets of 15 art prints reproducing paintings by Claude Monet or Aboriginal paintings and ranked them according to their preferences. Then, six groups of two pairs of prints were presented and subjects indicated for each group which pair they preferred. Five of these groups included novel images, while the last one contained prints that had already been presented. This group consisted of the pair of 4th- and 10th- ranked prints and the pair of the 6th- and the 12th- ranked prints from the same set. In the third phase, participants were asked to re-rank all the prints according to their current preferences. Finally, memory of the participants was assessed by asking them to identify the 4 prints that were presented both in the raking and the choice phases and to state for each print if it had been chosen or rejected. In this experiment, both amnesic patients and control subjects showed a significant spread. Interestingly, control participants were significantly better at identifying which prints they chose or rejected, compared to amnesic

patients, although the performance in the identification of the four critical prints was similar in both groups. In this study, the explicit memory of the choice does not seem critical for exhibiting cognitive dissonance reduction. Note, however, that this study is affected by the artefact highlighted by Chen and Risen (2010). Indeed, the authors used the second set of prints as a control condition, in order to assess the baseline effect of preference change, but they considered the pairs as selected or rejected based on the choice that was made on the other set of prints, which does not reveal any information about the actual preferences for the pairs of the non-critical set. The results are still interesting, because they address the issue of the cognitive processes involved in cognitive dissonance resolution, but they are not really conclusive as they are not controlled for the Chen and Risen's artefact. It would therefore be interesting to replicate these results in a properly controlled design, to assess the role of explicit memory.

Another study suggests that the role of explicit memory is not critical for cognitive dissonance resolution (Coppin et al., 2010). Participants were presented with twelve odorants and asked to rate them. Based on these ratings, pairs of odorants were created for the choice phase. Participants were presented with six odor pairs, including two pairs of similarly rated odors (difficult choices) and two pairs of odors rated differently (easy choices) and they were asked to choose the odor they preferred in each pair. Then, subjects were asked to rate again the pleasantness of the twelve odors. Finally, participants were presented again with the twelve odors together with six new odors and they were asked whether they had already smelled each odor and if yes, whether they had chosen or rejected it. The authors found significant choice-induced preference change in difficult choices, but not in easy choices, similarly to previous studies (Brehm, 1956; Sharot et al., 2009). Interestingly, when they analyzed the effect of memory, they found that spreading of alternatives could be observed both for remembered and forgotten choices (Fig. 1.13). Note, however, that in this study, the artefact highlighted by Chen and Risen (2010) is not corrected, casting doubt on the empirical results, although the question of understanding the processes underlying cognitive dissonance effect is very interesting.

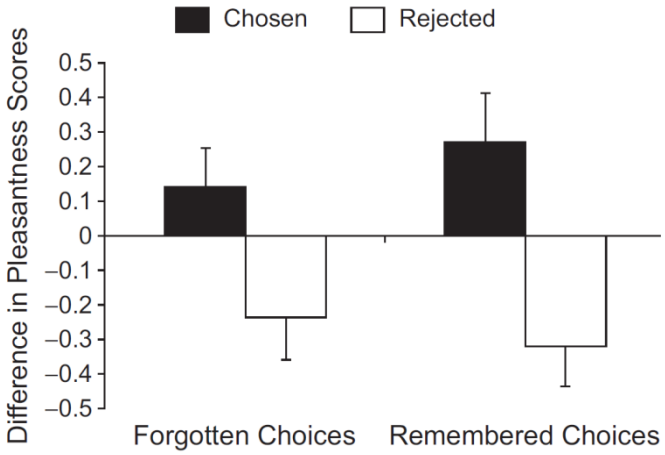

FIGURE 1.13: LEARNING OF SEQUENCES

Spreading of alternatives was observed for both forgotten and remembered choices. Adapted from Coppin (2010).

Finally, another study indirectly suggested that the memory of the choice is not critical for preference change to occur (Sharot et al., 2012). In this experiment, the authors tested whether choice-induced preference changes are long lasting. Participants were presented with 80 names of vacation destinations and were asked to rate them according to their liking of these destinations. Then subjects were asked to choose between pairs of vacation destinations (70% difficult choices, 30% easy choices) in the experimental condition or to indicate which destination the computer had randomly chosen in the control condition. Then participants in both groups rated the destinations. In the experimental condition, subjects had to provide a third rating approximately 3 years after the first phase of the experiment. In the control condition, participants provided a third rating 2.5 years of the first phase and completed a post-experimental free-choice decision about 6 months after (so approximately 3 years after the initial experimental phase). This control condition corresponds to the "rate-rate-choose" condition suggested by Chen and Risen (2010). This experimental design was similar to the one used by Izuma et al. (2010), although the experimental and control conditions were assigned in a within-subject design in this last study. Sharot et al (2012) found that choice-induced preference change can be observed in the experimental group, both when taking into account the

immediate and the delayed rating after the choice. These results suggests that explicit memory of the choice is probably not necessary to give rise to cognitive dissonance reduction, as it is highly unlikely that subjects would remember the choices they made in a psychology experiment three years ago. However, some methodological problems in this study call to a cautious interpretation of these results. First, the authors try to address the issue raised by Chen and Risen (2010), by including a "rate-rate-choose" condition. But the choice in this condition is made 3 years after the initial phase of the experiment, so one can make the reasonable assumption that subjects' choices could be different after 3 years from the choices they would have made immediately. This casts doubt on the validity of this choice to reveal underlying preferences at the time of the first rating. Second, while there was a significant difference between the "rate-choose-rate" and the "rate-rate-choose" conditions in the immediate preference change results, no such difference exists for the long-lasting preference change.

This series of three experiments tends to suggest that explicit memory of choices is not necessary to observe a spreading of alternatives in the free-choice paradigm. However, given the methodological issues raised above, these results should be confirmed with properly designed experiments which control for the Chen and Risen's artefact (Izuma and Murayama, 2013).

In the experimental studies presented in this PhD, I address this issue of the role of memory in cognitive dissonance resolution (Chapter 4 and Chapter 5). We designed a free-choice paradigm including a "rate-rate-choose" condition and assessed the memory of the choice at the end of the experiment. We explored the behavioral spreading of alternatives when choices were remembered or forgotten (Chapter 4) and we analyzed the activity in cerebral regions associated with episodic memory in this paradigm (Chapter 5).

3.3. Summary

- When confronted with inconsistencies in their own behavior, subjects tend to update their own values and attitudes.
- The cognitive dissonance theory provides a framework explaining these changes.
- Several experimental paradigms have been used to test the cognitive dissonance theory, but we focus on the free-choice paradigm in this PhD work.
- A major statistical artefact has been identified in the free-choice paradigm, but it can be corrected by appropriate control conditions.
- Cognitive processes underlying cognitive dissonance effects remain unclear. In particular, the role of explicit memory is disputed.
- Classic theories implicitly suggest that memory is necessary for cognitive dissonance to arise, whereas recent experimental studies challenge this view.

4. Presentation of the experimental work

In this introduction, I presented different examples both in patients and in control subjects suggesting that our interpretations of our experiences tend to maximize consistency and meaningfulness. These interpretations can be related to the integration of priors about the world with data sampled in the environment. The aim of this PhD thesis is to better understand how subjects change their interpretations when facing inconsistencies (both in the environment and in their own behavior) and how this process interacts with conscious access.

To address this issue, we conducted four studies:

The first study (Chapter 2) explored the neural bases of how subjects update their priors based on regularities in the environment and how they process information contradicting these priors. The second aim of this study was to understand what dissociates automatic regularity detection from a more controlled one. Using two levels of auditory regularities, we studied the neural mechanisms of automatic and strategic novelty detection. We took advantage of the high spatial and temporal resolution of the intracranial EEG to compare the dynamics of these two processes.

The second study (Chapter 3) aimed at understanding how subjects adapt their behavior based on their updated priors about the environment after being exposed to regularities and how stimuli which are inconsistent with these new priors influence subjects' responses. This second study also allowed us to test the existence of adaptation when the stimuli remained non-conscious. We designed an experimental paradigm derived from the Stroop task, in which 80% of trials were incongruent. In this condition of high conflict frequency, behavioral measures and EEG responses were collected to assess adaptation to conflict. Moreover, conscious and non-conscious trials were compared to test the existence of conflict adaptation in the absence of awareness.

The third study (Chapter 4) was designed to test how subjects deal with inconsistencies in their own behavior. We studied this issue in the context of the cognitive dissonance theory and used the free-choice paradigm. We corrected for the critical statistical artefact highlighted by Chen and Risen (2010) and identified a crucial role of memory in choice-induced preference change.

To further explore this process, we conducted the fourth study (Chapter 5) which examined the neural correlates of choice-induced preference change in a properly controlled setting. Moreover, this study confirmed the role of memory. Finally, it allowed us to test whether preference change occurred right after the choice or whether it was induced by the second rating.

CHAPTER 2. EVENT-RELATED POTENTIAL, TIME-FREQUENCY AND FUNCTIONAL CONNECTIVITY FACETS OF LOCAL AND GLOBAL AUDITORY NOVELTY PROCESSING: AN INTRACRANIAL STUDY IN HUMANS.

1. Presentation of the article

In this first experimental study, we explored the way subjects deal with inconsistencies in the environment. We used an experimental paradigm in which two levels of auditory regularities were embedded: a local regularity defined by the identity of sounds within series of five sounds and a global regularity defined across series of five sounds on a longer timescale. These two types of regularities are known to elicit different processes of regularity detection. It has been shown that the first one is automatic, while the second one is more strategic. We took advantage of the high spatial and temporal resolution of intracranial EEG to explore the dynamics of these two processes. We analyzed event-related potentials, time-frequency representations and functional connectivity in response to local and global deviant stimuli. This study allows us to assess how subjects identify different types of regularities in their environment and how they process stimuli inconsistent with these regularities.

2. Abstract

Auditory novelty detection has been associated with different cognitive processes. Bekinschtein et al (2009) developed an experimental paradigm to dissociate these processes, using local and global novelty, which were associated respectively with automatic versus strategic perceptual processing. They have mostly been studied using event-related potentials (ERPs), but local spiking activity as indexed by gamma (60-120 Hz) power and interactions between brain regions as indexed by modulations in beta band (13-25 Hz) power and functional connectivity have not been explored. We thus recorded nine epileptic patients with intracranial electrodes to compare the precise dynamics of the responses to local and global novelty. Local novelty triggered an early response observed as an intracranial mismatch negativity (MMN) contemporary with a strong power increase in the gamma band and an increase in connectivity in the beta band. Importantly all these responses were strictly confined to the temporal auditory cortex. By contrast, global novelty gave rise to a late ERP response distributed across brain areas, contemporary with a sustained power decrease in the beta band (13-25 Hz) and an increase in connectivity in the alpha band (8-13 Hz) within the frontal lobe. We discuss these multi-facet signatures in terms of conscious access to perceptual information.

Keywords: intracranial recordings, mismatch negativity, P300, time-frequency, connectivity

3. Introduction

Novelty detection is fundamental to quickly respond to potentially relevant stimuli. It is notably important to detect both unusual objects (perceptual novelty), and usual objects delivered in in a new context. This second type of novelty, termed 'contextual novelty', has been widely studied using event-related potentials (ERPs). Two main cognitive processes, a fast and automatic novelty detection process and a slow and strategic one, have been identified. The first one is indexed by the mismatch negativity (MMN - Näätänen et al., 1978), which is elicited around 100-200 ms after change onset. The MMN is generated by bilateral sources in the superior temporal cortex and possibly in the frontal cortex (Giard et al., 1990; Rinne et al., 2000; Opitz et al., 2002). These generators were confirmed by intracranial studies conducted in humans (Halgren et al., 1995a; Kropotov et al., 2000; Liasis et al., 2001; Rosburg et al., 2005). Interestingly, the MMN persists in the absence of attention (Näätänen et al., 1978). It can even be observed during rapid eye-movement sleep (Atienza et al., 1997) or in comatose patients (Fischer et al., 1999; Naccache et al., 2005) or more generally in patients with impaired consciousness (Faugeras et al., 2011, 2012; King et al., 2013). Following the MMN, another component sensitive to auditory novelty, the P300, can be recorded about 300 ms after change onset (Sutton et al., 1965) and is related to a strategic novelty detection process. Two types of P300 can be distinguished. The P3a is anteriorly distributed, peaks between 220 and 280 ms and is weakly affected by attention (Squires et al., 1975), whereas the P3b is a centro-posterior component, reflecting the activation of the hippocampus and parietal and frontal cortices (Picton, 1992), and spanning from 250 to 600ms after change onset. The P3b is highly dependent on attention, as it only occurs when subjects are actively engaged in detecting novel stimuli (Bekinschtein et al., 2009). Therefore, it has been proposed as a neural signature of working memory (Donchin and Coles, 1988; Polich, 2007) and conscious access (Sergent et al., 2005).

Bekinschtein et al (2009) developed an experimental paradigm exploring these two processes. It relies on two levels of auditory novelty which are orthogonally manipulated: a change in pitch

within series of five sounds (local novelty) gives rise to an MMN, whereas a rare change in series of five sounds in a fixed context (global novelty) triggers a P3b. An fMRI experiment using this exact paradigm showed that local novelty elicited responses within auditory cortices, whereas processing of global novelty was associated with a distributed brain network including frontal, anterior cingulate and parietal areas (Bekinschtein et al., 2009). Interestingly, the response to local novelty was present even when subjects were distracted, whereas the response to global novelty only occurs when subjects were actively engaged in detecting novel stimuli (Bekinschtein et al., 2009) or when they were paying attention to the stimuli (Wacongne et al., 2011). Moreover, using this paradigm, responses to local novelty were observed in patients with impaired consciousness, but responses to global novelty were only observed in patients with preserved consciousness (Faugeras et al., 2011, 2012; King et al., 2013).

Taken together, these findings suggest that local novelty detection involves a fast, automatic and encapsulated process, whereas global novelty detection involves a slow, strategic and widespread process.

Nevertheless, responses to local and global novelty have mostly been explored using ERP measures, but it is important to explore their oscillatory properties and functional connectivity in order to fully understand the networks underlying these two types of responses, as evidenced in previous intracranial studies (Axmacher et al., 2010; Zaehle et al., 2013). Indeed, responses to events of interest contains modulations in some frequency bands that are time-locked to the events, but not phase-locked, and thus cannot be extracted by ERPs (Pfurtscheller and Lopes da Silva, 1999; Tallon-Baudry and Bertrand, 1999). In particular, modulations in high gamma activity has been proposed to reflect local spiking activity (Pesaran et al., 2002; Nir et al., 2007; Ray and Maunsell, 2010), which cannot be measured by ERPs. Moreover, interactions between brain regions cannot be explored by ERPs but are thought to be reflected in modulations of power in the beta band (Kopell et al., 2000) and in functional connectivity measures, such as phase coupling(Fries, 2005).

We thus compared spatio-temporal dynamics of the responses to local and global novelty by analyzing ERPs, spectral power and functional connectivity, in nine epileptic patients with intracranial electrodes, taking advantage of the high spatial and temporal resolutions of these recordings.

We predicted responses to local novelty to be confined within superior temporal cortices and to appear as an intracranial MMN ERP, contemporary with an increase in gamma-band power indexing localized neural activity (Pesaran et al., 2002) and associated with no clear increase in long-range functional connectivity. In sharp contrast, we predicted responses to global novelty to be more widespread across brain regions, including in particular parietal and frontal areas: indeed global novelty detection involves working memory, and fronto-parietal areas have been associated to this effect in MEG (Wacongne et al., 2011) and fMRI (Bekinschtein et al., 2009) studies. We also predict that these responses should appear as a maintained ERP peaking around 300ms, contemporary with a modulation in beta band power which has been associated with long-range communication (Donner and Siegel, 2011), and with direct evidence of inter-area connectivity.

4. Materials and Methods

4.1. Patients

Ten epileptic patients (age M = 32 year old, SD = 11 years; 4 males – see Table 1) gave their written informed consent to participate in this study. Neuropsychological assessment revealed normal or mildly impaired general cognitive functioning (IQ ranged from 77 to 98) for all patients but one. This last patient had an extremely low IQ of 64, but was able to do the task properly (97% accuracy). These patients suffered from drug-refractory focal epilepsy and were implanted stereotactically with depth electrodes as part of a presurgical evaluation. Implantation sites were selected on purely clinical criteria, with no reference to the present protocol. This experiment was approved by the ethical committee of Pitié-Salpétrière Hospital (Comité Consultatif de Protection des Personnes participant à une Recherche Biomédicale). Based on poor behavioral performance (only 56% accuracy), patient 10 was excluded.

Patient	Age	Gender	Handedness	Epilepsy duration	Total Number of Electrodes	Temporal Electrodes	Frontal Electrodes	Occipital Electrodes
1	18	F	R	8 years	36	36	0	0
2	29	F	R	4 years	25	25	0	0
3	48	F	L	10 years	37	37	0	0
4	46	F	R	25 years	14	11	3	0
5	23	M	R	8 years	43	10	33	0
6	43	F	R	24 years	27	11	16	0
7	26	M	L	16 years	42	40	0	2
8	42	M	R	26 years	9	0	9	0
9	26	M	R	11 years	49	24	0	25
10	*24*	*F*	*R*	*14 years*	*72*	*9*	*63*	*0*

TABLE 1: CLINICAL CHARACTERISTICS OF THE PATIENT SAMPLE

4.2. Procedure

The procedure used in this experiment exactly followed the "Local-Global" paradigm, developed by Bekinschtein et al. (2009), who already reported intracranial ERPs from two of the present patients.

Patients were presented with 8 blocks of 123 to 128 trials. In each trial, a series of five sounds was played over a total of 650 ms. The first four sounds were always identical, either low (sound

A) or high-pitched (sound B), but the fifth one could be either identical (AAAAA or BBBBB) or different (AAAAB or BBBBA). Thus this last sound respects or violates the local regularity established by the first four sounds in local standard or local deviant trials, respectively (Fig. 2.1). On top of this local rule, a global regularity was added. In each block, global standard trials were delivered on 80% of trials. In the first type of blocks, these trials were local standard trials, whereas in the second type of blocks, they were local deviant trials. By contrast, global deviant trials were presented in 20% of trials in a given block (Fig. 2.1). In the first type of blocks, these trials were local deviant trials, whereas in the second type of blocks, they were local standard trials. The first 20 to 30 trials of each block were global standard trials to establish the global regularity. Local and global regularities were thus manipulated orthogonally (Fig. 2.1).

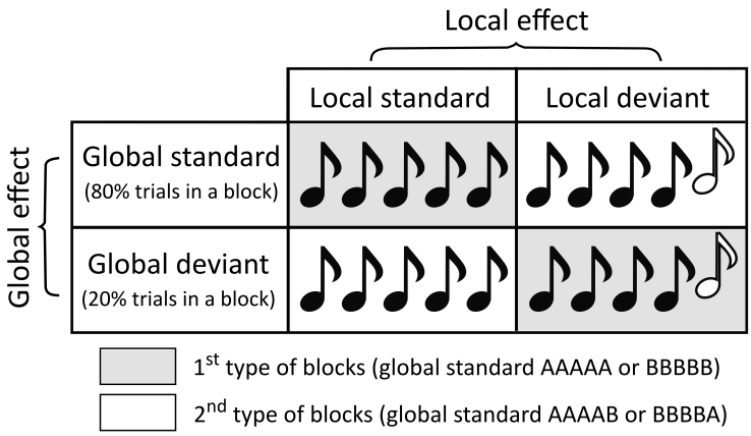

FIGURE 2.1: THE LOCAL GLOBAL PARADIGM

The Local Global paradigm (Bekinschtein et al., 2009) is an auditory oddball paradigm with two levels of regularity. Each trial is composed of a series of five successive sounds (SOA = 150 ms). The first four sounds are always identical. The fifth sound can either be identical to these first sounds in local standard trials or different in local deviant trials. On top of this local regularity, a global rule is added. Global deviant trials correspond to a series of five sounds which is rare in a given block, compared to the frequent global standard trials. Local and global regularities are manipulated orthogonally, resulting in four types of trials: local standard-global standard and local deviant-global deviant in the first type of blocks, and local deviant-global standard and local standard-global deviant in the second type of blocks. Two sounds were used to generate these trials: sound A (composed of 350, 700 and 1400 Hz sinusoidal tones) and sound B (composed of 500, 1000 and 2000 Hz sinusoidal tones). The local effect corresponds to the difference between local deviant and local standard stimuli, whereas the global effect corresponds to the difference between global deviant and global standard stimuli.

This design enables the comparison between physically identical stimuli in different contexts. In the following analyses, we will refer to local and global effects, which correspond respectively to the contrast between local deviant vs. local standard stimuli, and between global deviant vs. global standard stimuli (Fig. 2.1). Additionally, patients were instructed to actively count the number of global deviant trials and report this number at the end of each block. This task ensured that they were paying attention to the stimuli.

Each sound was 50-ms long and composed of 3 sinusoidal tones (350 Hz, 700 Hz and 1400 Hz, sound A; or 500 Hz, 1000 Hz and 2000 Hz, sound B). All tones were prepared with 7-ms rise and 7-ms fall times. Four different series were used: AAAAA, BBBBB, AAAAB and BBBBA. The Stimulus Onset Asynchrony (SOA) between sounds was 150 ms. Series of sounds were separated by a variable silent interval of 1350 to 1650 ms. Four different blocks were thus created, with each possible series of sounds as global standard trials. All patients heard each of these blocks twice, in randomized order. Sounds were presented from the computer's speakers, using E-prime v1.2 (Psychology Software Tools Inc.) at the bedside of each patient.

4.3. Electrode implantation and localization

Patients were implanted intracerebrally with depth electrodes, each bearing four to ten recording sites (Ad-TechMedical Instruments, Racine, WI, US). Each patient had on average 59 (SD = 13) recording sites.

To compare recording sites position and summarize brain activations across patients, their coordinates were obtained after normalizing the anatomical three-dimensional post-implantation MRI onto the template from the Montreal Neurological Institute (MNI), using SPM8 software (http://www.fil.ion.ucl.ac.uk/spm). Electrodes localization plots used the iso2mesh toolbox (Fang and Boas, 2009; Fang, 2010).

4.4. Data acquisition and preprocessing

For seven patients, data were acquired with an audio-video-EEG monitoring system (Micromed, Treviso, Italy), which allowed for simultaneous acquisition of data from up to 128 EEG channels

sampled at 1024 Hz. Data from the remaining three patients were acquired using another audio-video-EEG monitoring system (Nicolet-Viasys, Madison, WI, USA), which allowed for simultaneous acquisition of data from up to 64 EEG channels sampled at 400 Hz.

Unless specified otherwise, data were analyzed with Fieldtrip toolbox (Oostenveld et al., 2011) and Matlab 2011a (The Mathworks, Inc., Natick, MA, USA). All the analyses were done at the electrode level, as the position of the recording sites differed from one patient to another.

Epochs were extracted (-800 ms to 700 ms after the onset of the fifth sound). To avoid artifacts, recording sites exceeding the threshold of ±300 μV in more than 5% of the epochs were excluded. All signals were re-referenced to their nearest neighbor on the same electrode (bipolar montage). In the following, we will refer to these bipolar montages as 'electrodes'. All data were visually inspected to discard any trial with epileptic activity.

4.5. Behavioral analysis

Patients were instructed to count silently the number of global deviants and report this number. For each patient, we computed the average percentage of errors over blocks and we report accuracy as 100% minus this percentage.

4.6. ERP analysis

Event related potentials (ERPs) were obtained by averaging epochs for each condition. The signal was band-pass filtered off-line from 0.5 to 20 Hz using a fourth-order Butterworth filter in forward and reverse directions in order to avoid phase-shift and a baseline correction was applied, by subtracting the mean voltage in the [-800 ms 0 ms] window.

To assess the statistical significance of our results, ERPs of local deviant trials and global deviant trials were compared to those of local standard trials and global standard trials, respectively, using independent sample t-tests for each electrode. To control for type I errors generated by multiple comparisons across time at level α, we used a nonparametric procedure. We computed N = 1000 permutations by shuffling trial labels. For each permutation, the maximal t-value

across time samples was extracted to estimate the permutation distribution of the maximal statistic. The critical threshold which controls for family-wise error rate over time samples was define as the c + 1 largest member of this distribution, where c is equal to αN rounded down. In the original data, only samples with a t-value higher than this threshold were identified as significant (Nichols and Holmes, 2002). In addition, to correct for multiple comparison over electrodes when visualizing effects across all electrodes in all patients, we used a False Discovery Rate (FDR) correction on p-value obtained at the electrode level, across the whole time window. All reported p-values are corrected and referred to as p_{corr}.

Peak latencies were identified within periods of statistical significance on the difference between standard and deviant trials. For the sustained responses to global novelty, latencies were estimated as the earliest significant differences given the absence of a clear peak. In both cases, reported latencies are the average across all significant electrodes.

4.7. Time-frequency analysis

Time-frequency representations were calculated by Morlet wavelets, as described previously (Tallon-Baudry et al., 1997). The power at a given time t and frequency f_0 is given by the squared norm of the convolution of the signal to the wavelet $w(t, f_0)$:

$$w(t, f_0) = A exp\left(\frac{-t^2}{2\sigma_t^2}\right) exp(2i\pi f_0 t) \quad where\ A = (\sigma_t \sqrt{\pi})^{-1/2}$$

The width of the wavelet $m = 2\pi\sigma_t f_0$ was set to 5 as this value gives a good tradeoff between time and frequency resolution (De Moortel et al., 2004). The length of each wavelet used for the computation was $3\sigma_t$. The convolution of the signal by this set of wavelets resulted in an estimate of power at each time sample and at each frequency (2 Hz step) between 5 and 200 Hz. Power was then converted to a decibel (dB) scale. Time-frequency values were obtained from the subtraction between two conditions, without baseline correction.

To assess the significance of differences in oscillation power across conditions, we used independent sample t-tests at each time and frequency point. Then correction for multiple

comparisons over time and frequency was performed, by nonparametric cluster-based method (Maris and Oostenveld, 2007) . We computed N = 1000 permutations by shuffling trial labels. Then, for each permutation, independent sample t-tests were performed at each time and frequency sample. All samples with a t-value corresponding to a p-value smaller than 0.05 were clustered in connected sets on the basis of adjacency in time and frequency. Then the cluster statistic was computed by taking the sum of the t-values within each cluster. The cluster-corrected threshold was obtained by computing the permutation distribution of the maximum cluster statistic and taking the c+1 largest member of this distribution, with c is equal to αN rounded down. In the original data, only clusters with a cluster statistic higher than this threshold were identified as significant. Moreover, when visualizing effects across all electrodes in all patients, we additionally corrected for multiple comparisons over electrodes, using False Discovery Rate (FDR) correction. All reported p-values are corrected and referred to as p_{corr}.

Peak latencies were identified within time-frequency windows of statistical significance on the difference between standard and deviant trials. For sustained responses, effect latencies were estimated as the earliest significant differences between deviant and standard trials. In both cases, reported latencies are the average latency across all significant electrodes.

4.8. Functional connectivity analysis

Connectivity analyses were performed using pairwise phase consistency (PPC – Vinck et al., 2010) which provides a method for measuring rhythmic synchronization, without being affected by the finite sample size bias, observed with classic tools, such as phase locking value (PLV – Lachaux et al., 1999). This issue is particularly relevant when comparing global deviant and standard conditions, as the number of trials in these conditions is different. In each trial j, for each time point t and frequency f, the phase $\varphi_j\ t, f$ of the signal was estimated, using wavelets, as presented above. Pairwise phase consistency PPC_{kl} between electrodes k and l is then computed across N trials as follows:

$$PPC_{kl}\ t\ = \frac{2}{N\ N-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} cos(\ \Delta\varphi_{kl}^{i}\ t\ -\ \Delta\varphi_{kl}^{j}\ t\ )$$

with $\Delta\varphi_{kl}^{i}\ t$ the phase difference between electrodes *k* and *l* in trial *i* at time *t*

Pairwise phase consistency was computed for all pairs of electrodes, for frequencies between 5 and 60 Hz, between – 800 and 700 ms relative to the onset of the fifth sound. No baseline correction was applied.

Statistical significance was assessed by a nonparametric test on the difference of pairwise phase consistency between conditions (Lachaux et al., 1999), using 500 permutations, for each time and frequency samples and cluster-based correction for multiple comparisons across time and frequency samples was applied (Maris and Oostenveld, 2007). All reported p-values are corrected and referred to as p_{corr}. Given the high level of correlations between pairs of electrodes and the relatively low number of highly computationally demanding permutations, multiple comparisons are corrected only across time and frequency samples, but not across pairs of electrodes in order to avoid missing potentially relevant changes in functional connectivity (type II errors). However this less conservative threshold could lead to false-positives (type I errors). This is why we report the average PPC changes across all pairs of temporal and frontal electrodes (see Fig. 2.4b).

The estimation of connectivity is based on the consistency of phase lags between areas. Note that consistent phase lags reflect the true connectivity between areas but also artifactual connectivity driven by synchronization of different areas onto any external reference. Therefore, we tried to isolate genuine synchronization between electrodes from spurious connectivity related to simultaneous processing of the stimuli (e.g. in left and right auditory cortices). We reasoned that spurious connectivity would be linked to oscillations evoked by the stimulations at several electrodes, as opposed to the induced activity which is time-locked, but not phase-locked, to the stimulation. We introduced a novel analysis, the induced PPC, based on the 'evoked' vs. 'induced' terminology introduced by Tallon-Baudry and Bertrand (1999). For this

analysis, the evoked oscillations were regressed out of the data and connectivity was estimated using the phase of the induced oscillations isolated in this way. Thus, induced PPC measures functional connectivity, which is not related to external stimulations. To compute this measure, we first averaged the data across trials in a given condition. This average preserves evoked oscillations and cancels out induced oscillations (Tallon-Baudry and Bertrand, 1999). We then computed the time-frequency decomposition of this average using wavelets as described above. To isolate the induced response, we decompose data into time and frequency and regressed out linearly from each trial of a given condition the time-frequency decomposition of the corresponding evoked response. This corrected signal was used to compute the $PPC_{induced}$ and statistical comparisons were performed as previously described for the PPC.

5. Results

Auditory novelty detection was studied using 282 intracranial electrodes across 9 epileptic patients, with an average of 31 (SD = 13) electrodes per patient (Table 1). Activity was recorded from the temporal lobe (194 electrodes), the frontal lobe (61 electrodes) and the occipital lobe (27 electrodes). Note that, as we did not have precise hypotheses about the laterality of the effects, we analyze electrodes from the left and right hemispheres together. All these patients performed the task properly (accuracy higher than 80%). We analyzed event-related potentials (ERPs), time-frequency decompositions and functional connectivity associated with local and global auditory novelty detection. This approach allowed us to test whether local novelty was related to an automatic and encapsulated process and whether global novelty implicated distinct inter-connected brain regions.

5.1. ERP analysis

Significant responses to violation of local regularity (local deviant compared to local standard trials) were observed in 44 electrodes (16% of all the electrodes, t-test, all $p_{corr}<0.05$), all located in the temporal lobe. These responses were observed in 6 out the 8 patients with temporal electrodes. Notably, 20 out of the 26 electrodes implanted in the superior temporal lobe (77%) showed significant local effect (Fig. 2.2a). The responses to local novelty presented two main components: an early one, with a peak on average at 133 ms (SD = 17 ms – see Fig. 2.2b for an exemplar electrode in which this peak is at 105 ms) after the onset of the fifth sound, and a second component, with a reversed polarity, showing a peak on average at 231 ms (SD = 33 ms – see Fig. 2.2b for an exemplar electrode in which this peak is at 165 ms). This local effect is related to the mismatch negativity observed in scalp EEG (Näätänen et al., 2001) and was found in previous intracranial studies (Halgren et al., 1995a; Edwards et al., 2005; Bekinschtein et al., 2009). Note that, interestingly, the polarity of this component depends on the position of the electrode relative to the Sylvian fissure (Halgren et al., 1995b). Moreover, in 8 of these 44 electrodes, from 3 different patients, these components were preceded by an event peaking on

average at 71 ms (SD = 8 ms – see Fig. 2.2b for exemplar electrode in which this peak is at 63 ms) and which can be related to early components of novelty detection found in scalp EEG (Grimm et al., 2011). Finally, in 18 electrodes from 4 patients, the two main components were followed by an event with a peak on average at 435 ms (SD = 95 ms – see Fig. 2.2b for exemplar electrode in which this peak is at 330 ms), which may be related to the P3a (Halgren et al., 1995a).

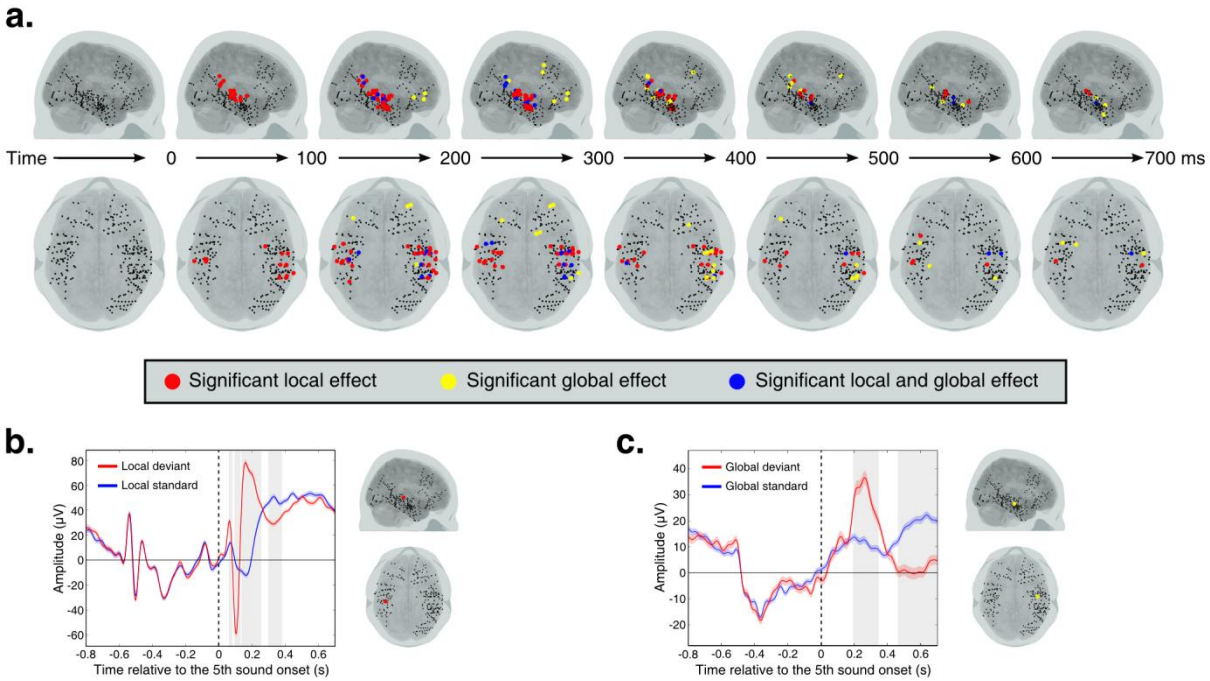

FIGURE 2.2: EVENT-RELATED POTENTIALS IN RESPONSE TO LOCAL AND GLOBAL NOVELTY

(a) Localization of significant differences between standard and deviant stimuli. Each black dot represents an electrode. Red, yellow and blue dots represent electrodes showing a significant local, global and both local and global effect respectively in the time window indicated at the extremities of the black arrows.

(b) Example of a temporal electrode showing significant differences between local deviant (in red) and local standard (in blue) stimuli. Four components, peaking at 63 ms, 105 ms, 165 ms and 330 ms, can be identified. Red and blue shadings represent SEM. Gray shading represents significant differences between conditions (p_{corr} < 0.05). The precise localization of this electrode is shown on the right.

(c) Example of a temporal electrode showing differences between global deviant (in red) and global standard (in blue) stimuli. Two components can be identified: the first one shows a peak at 260 ms and the second one starts at 462 ms. Red and blue shadings represent SEM. Gray shading represents significant differences between conditions (p_{corr} < 0.05). The precise localization of this electrode is shown on the right.

By contrast, responses to violation of global regularity (global deviant compared to global standard trials) were observed in 32 electrodes (11% of all the electrodes). These responses were more widespread across cortical structures than the local effect and were observed in 7 out 9 patients. Twenty-four electrodes in the temporal lobe (24% of temporal electrodes), including 12 in the superior temporal lobe (46% of superior temporal lobe electrodes), and 8 in the frontal lobe (13% of frontal electrodes) showed significant effects (t-test, all $p_{corr} < 0.05$ – Fig. 2.2a). The global effect was associated with three distinct components. First, in 19 electrodes from 5 patients, we observed a transient difference between global deviant and global standard trials on average at 226 ms (SD = 56 ms – see Fig. 2.2c for an exemplar electrode in which this component peaks at 260 ms), similar to the second component observed in response to local novelty. Indeed, only temporal electrodes showing this local effect presented such an early global effect. This component could be explained by a contextual modulation of the MMN amplitude. Indeed, local deviant stimuli are used in our paradigm both as global standards and as global deviants (Fig. 2.1). Thus, the probability of local deviant stimuli is not the same in each block and this could affect the amplitude of the MMN (Sato et al., 2000; Wacongne et al., 2011; King et al., 2013). Second, a sustained difference starting on average at 366 ms (SD = 131 ms – see Fig. 2.2c for an exemplar electrode in which this component starts at 462 ms) was observed in response to global novelty in 20 electrodes from 7 patients. It has been associated to the P3b response observed in scalp EEG (Bekinschtein et al., 2009) and has been reported in previous intracranial studies (Smith et al., 1990; Baudena et al., 1995; Halgren et al., 1998). Note that the polarity of this component depends on the localization relative to the different generators (Smith et al., 1990). Finally, two electrodes from one patient implanted in the anterior cingulate cortex showed a transient response, with a peak on average at 265 ms, but did not show any significant local effect. Note that we did not observe significant local or global effect in the occipital lobe. This may partly be explained by the low number of electrodes implanted in this region (only 27 electrodes, among which 25 belong to the same patient).

Note that the choice of the baseline time window could potentially influence the results, especially for the global effect as expectation related components can build up (Faugeras et al., 2012). However, the number of electrodes showing significant local and global effects was not different when choosing a baseline expanding across the whole trial (for local effect: $\chi^2 = 0.12$, $p=0.73$; for global effect: $\chi^2 = 0.69$, $p=0.41$).We thus identified different ERP responses to local and global novelty. Local novelty was mainly associated with an early response restricted to the temporal local. By contrast, global novelty was associated with late responses in different brain regions.

5.2. Time-frequency analysis

Task-induced activity was then analyzed in the time-frequency domain, using Morlet's wavelets. Violation of local regularity induced an increase in power in the gamma band (60-120 Hz) peaking on average 135 ms (SD = 45 ms) after the onset of the fifth sound, as revealed by the analysis of the proportion of electrodes showing a significant effect ($p_{corr} < 0.05$) at each time-frequency bin (Fig. 2.3a for exemplar electrode and Fig. 2.3b). This effect was observed in 22 electrodes (8% of all the electrodes), mostly implanted in the temporal lobe, in 6 out the 9 patients. Among these electrodes, nine were located in the superior temporal lobe (35% of superior temporal electrodes – Fig. 2.3c). One frontal electrode showed a similar effect (2% of frontal electrodes).

Violation of global regularity gave rise to an early increase in the gamma band peaking at 182 ms (SD = 50 ms). This modulation in the gamma band was similar to the one observed in the local effect, but more sustained in time. The increase in gamma power was observed in 6 out of 9 patients, in 23 electrodes (8% of all the electrodes), mostly implanted in the temporal lobe as in the local effect, but it was also present in 5 frontal electrodes (8% of frontal electrodes), which, as reported previously, did not show a significant local effect (Fig. 2.3c). This first effect was followed by a decrease in the beta band (13-25 Hz) which started on average 258 ms (SD = 115 ms) after the fifth sound onset and was maintained until the end of the trial (Fig. 2.3a for

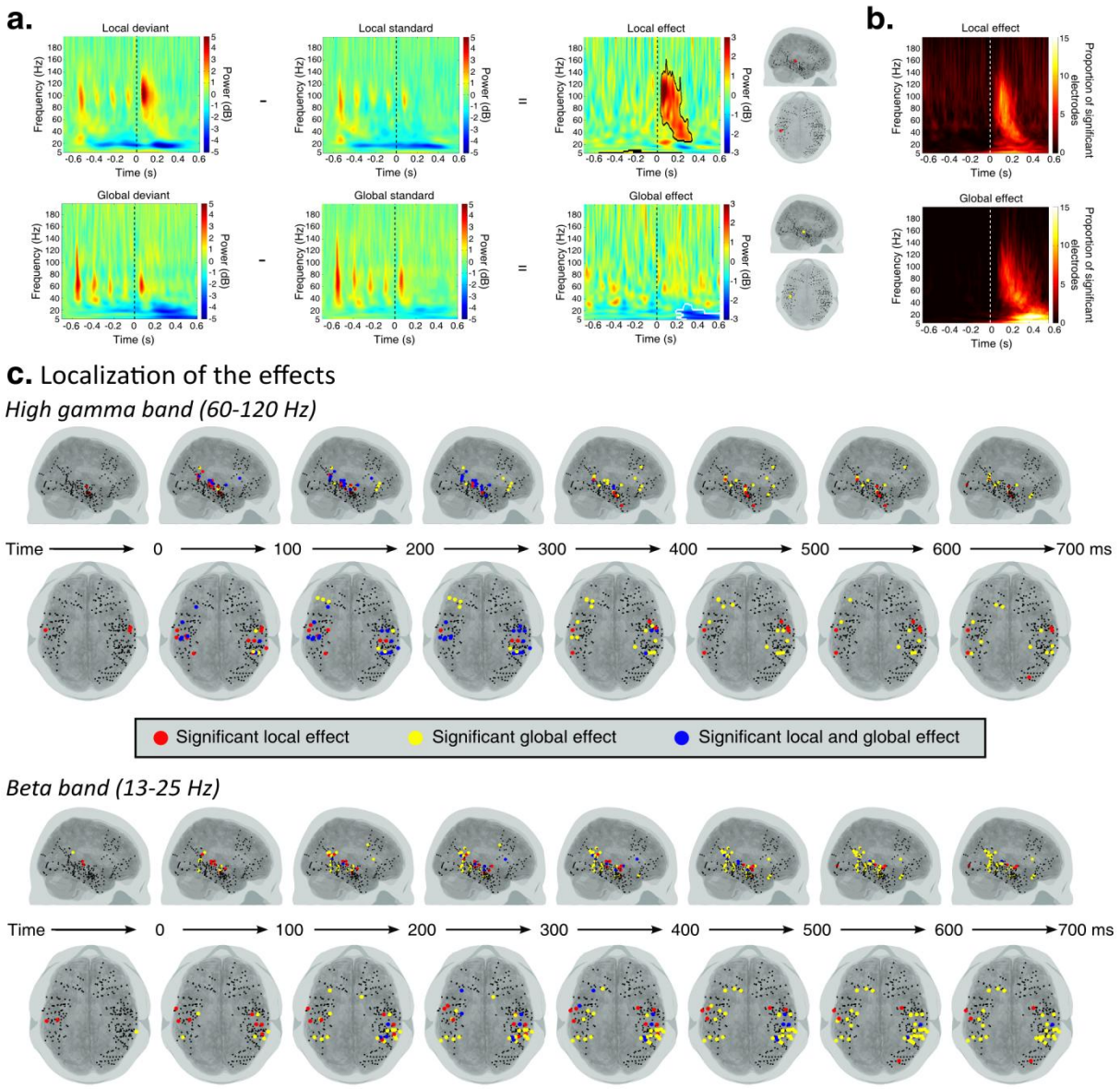

FIGURE 2.3: TIME-FREQUENCY ANALYSIS

(a) Example of electrodes showing significant local and global effect. Top row, time frequency responses of a temporal electrode to local deviant (left), local standard (middle) and local effect (difference between local deviant and local standard – right). Bottom row, time frequency response of a temporal electrode to global deviant (left), global standard (middle) and global effect (difference between global deviant and global standard – right). Black and white contours circle significant time-frequency samples ($p_{corr} < 0.05$). The precise localization of these electrodes is shown on far right.

(b) Proportion of significant electrodes ($p_{corr} < 0.05$) in each time and frequency bin for the local effect (top) and the global effect (bottom).

(c) Localization of significant effects in the high gamma band (top) and the beta band (bottom). Each black dot represents an electrode. Red, yellow and blue dots represent electrodes showing significant local, global and both local and global effects respectively in the time window indicated at the extremities of the black arrows.

exemplar electrodes and Fig. 2.3b). This decrease in beta band power was observed in 26 electrodes (9% of all the electrodes) from 6 out 9 patients and was widespread, affecting temporal but also 4 frontal electrodes (6% of frontal electrodes – see Fig. 2.3c). Similar to the ERP results, this beta band response, - which was sustained over time -, was only observed in response to global novelty. Interestingly, 16 of the 22 electrodes showing an increase in gamma band power in response to local novelty are also among the 44 electrodes showing a significant local ERP effect. Note that the latency of the gamma band response is very similar to that of the early response to local novelty observed in ERP. Moreover, 15 of the 23 electrodes showing an increase in gamma band power in response to global novelty are also among the 32 electrodes showing a significant ERP response to global novelty. Finally, only 7 electrodes belonged both to the set of 32 electrodes showing a significant ERP response to global novelty and to the set of 26 electrodes showing a decrease in the beta band in response to global novelty. These results highlight the fact that ERP and time-frequency analyses extract different facets of the responses to local and global novelty, as significant effects identified with both methods are not necessarily present in the same electrodes.

Local and global novelties were associated with different time-frequency responses. Violations of local regularity were related to an early increase in high gamma power. By contrast, violations of global regularity were associated with a sustained decrease in beta power. Interestingly, these time-frequency responses highlight different aspects of brain responses to local and global novelty than ERP responses.

5.3. Functional connectivity analysis

Changes in connectivity related to the task were analyzed using pairwise phase consistency (PPC), a measure of synchronization which is not biased by the number of trials in each condition (Vinck et al., 2010). We report results for all pairs of electrodes (n = 4714) to identify modulation of connectivity in response to local and global novelty, with a special focus on the temporal and frontal lobes, as most of the electrodes were implanted in these regions and no

significant response was observed in the other lobes in ERP and time-frequency analyses. There were 2507 pairs between temporal electrodes, 687 pairs between frontal electrodes, 539 pairs between temporal and frontal electrodes, 301 pairs between occipital electrodes and 680 pairs between occipital and temporal electrodes.

Local novelty was associated with a transient increase in functional connectivity, as measured by PPC, in the beta band (13-25 Hz), centered at 80 ms after the onset of the fifth sound (see Fig. 2.4a for exemplar pair). This increase was significant ($p_{corr} < 0.05$) in 38 pairs of electrodes (1% of all the electrode pairs), in 7 out of 9 patients. All of these pairs, but five, were between temporal electrodes (see Fig. 2.4b for the average across all pairs of temporal electrodes).

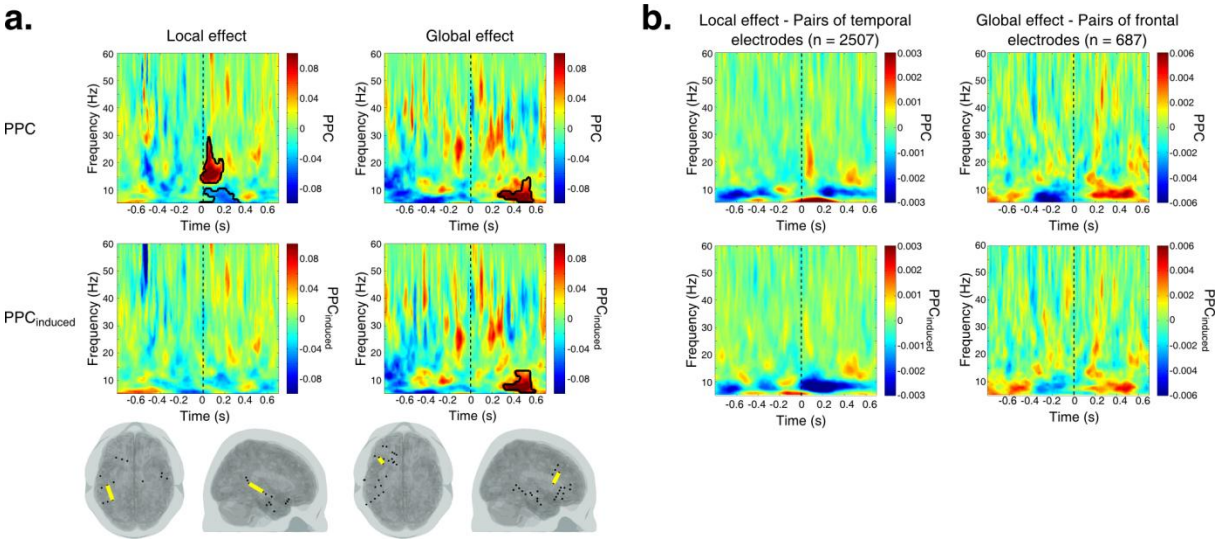

FIGURE 2.4: FUNCTIONAL CONNECTIVITY

(a) Time frequency representations of pairwise phase consistency (PPC – top row) and induced PPC (middle row) modulations in two pairs of electrodes. On the left, differences in PPC and induced PPC between local deviant and local standard stimuli in a pair of temporal electrodes are represented. On the right, differences in PPC and induced PPC between global deviant and global standard stimuli in a pair of frontal electrodes are represented. The bottom row shows the localization of these pairs. Black contours circle significant time-frequency samples ($p_{corr} < 0.05$).

(b) On the left, average differences in PPC (top) and induced PPC (bottom) between local standard and local deviant stimuli in pairs of temporal electrodes (n = 2507) are presented. On the right, average differences in PPC (top) and induced PPC (bottom) between global standard and global deviant stimuli in pairs of frontal electrodes (n = 687) are presented.

Interestingly, only 9 of them showed a similar increase in the PPC induced analysis. This suggests that the effect is related to activity evoked by the stimuli and can be partly explained by a parallel and simultaneous processing of the sounds rather than by a genuine inter-area exchange of information. Local novelty was also associated with a decrease in PPC in the beta band in 20 other pairs of temporal electrodes. Among them, 18 pairs showed a similar decrease in induced PPC.

On the other hand, global novelty was associated with a sustained increase in PPC in the alpha band (8-13 Hz), starting at 160 ms after the onset of the fifth sound and lasting until 600 ms after the onset of the fifth sound (see Fig. 2.4a for exemplar pair and Fig. 2.4b). This effect was observed in 7 out of 9 patients in 49 pairs of electrodes (1% of all electrode pairs): 27 between temporal electrodes (1% of temporal pairs), 7 between frontal electrodes (1% of frontal pairs), 5 between frontal and temporal electrodes (1% of temporo-frontal pairs) and 10 between occipital and temporal electrodes (1% of occipito-temporal pairs). Interestingly, among these electrode pairs, 42 showed this late increase in connectivity in the $PPC_{induced}$ analysis, which suggests that it is related to a genuine synchronization, and not simply to a simultaneous processing of the stimuli. Global novelty was also associated with a sustained decrease in PPC in 100 pairs of electrodes, in the same time window as the increase described above. This decrease was also observed in these pairs in the $PPC_{induced}$ analysis. Among these pairs, 66 were between temporal electrodes, 28 between occipital and temporal electrodes, 1 between temporal and frontal electrodes, 5 between occipital electrodes, 1 between frontal electrodes. Note, however, that this decrease in connectivity was mainly observed in patient 9, for whom 78 significant pairs were observed. This patient was the only one with occipital and posterior temporal electrodes.

Interestingly, the increase in PPC in response to local novelty implicated 45 electrodes, among which 16 also showed en ERP effect and 11 showed a time-frequency response to local novelty. The increase in $PPC_{induced}$ in response to local novelty implicated 19 of these 45 electrodes,

among which 7 showed an ERP effect and 5 showed an increase in gamma power in response to local novelty. In response to global novelty, the increase in PPC implicated 67 electrodes, among which 9 showed an ERP effect and 16 showed a time-frequency response to global novelty. Only 6 of these 67 electrodes were not in pairs of electrodes showing an increase in $PPC_{induced}$ in response to global novelty. Among these 6 electrodes, 3 showed a significant ERP response and 2 showed a time-frequency response to global novelty.

Local novelty was thus associated with an early increase in functional connectivity in the beta band, mostly in pairs of temporal electrodes. This increase was not present in most of these pairs in the $PPC_{induced}$ analysis, so it may probably partly explained by the simultaneous processing of the stimuli by different brain regions. By contrast, global novelty was associated with a late increase in alpha band functional connectivity. This increase was observed even when using $PPC_{induced}$, suggesting that this increase is related to genuine synchronization between brain areas.

6. Discussion

We studied the responses to two embedded levels of auditory novelty, defined at local (within trials) and global (between trials) scales, in nine epileptic patients implanted with 282 depth electrodes. We report for the first time a systematic description of processes underlying local and global novelty detection combining ERPs, time-frequency and functional connectivity measures. These three facets converged to delineate several differences between these two neural events, supporting the hypotheses that the local novelty detection is associated with an early process confined to the temporal lobe, while by contrast global novelty detection is processed later and implies the coordinated activity of distributed brain regions interacting together in the slow frequency range.

6.1. Two distinct neural events

Detection of local novelty was reflected in: i) two successive ERP components (peaking on average at 133 ms and 231 ms). This intracranial ERP response has previously been associated with the MMN observed in scalp EEG in response to local novelty, as it is observed in brain regions identified by source reconstruction of the MMN and it has a similar time course (Bekinschtein et al., 2009). It also has been reported in previous intracranial studies (Halgren et al., 1995a; Liasis et al., 2001; Edwards et al., 2005; Rosburg et al., 2005). Detection of local novelty was also reflected in ii) a transient increase in high gamma (60-120 Hz) activity, peaking on average at 135 ms after the onset of the critical sound, and iii) an increase in local connectivity, as measured by pairwise phase consistency (PPC), in the beta band (13-25 Hz). This increase was centered on average around 80 ms after the onset of the critical sound and was observed only in pairs of temporal electrodes. Interestingly, this change in connectivity could not be measured by the induced PPC in most of these pairs of electrodes, which suggests that it can be partially explained by a simultaneous and parallel processing of the stimuli in different recording sites (e.g. in the auditory cortex of both hemispheres). Importantly, those

three responses were exclusively observed within the temporal lobe, and more precisely in the superior temporal plane.

On the contrary, responses to global novelty were distributed in multiple cortical areas, and in particular in temporal and frontal regions. They comprised: i) a sustained ERP difference starting on average 366 ms after the onset of the fifth sound. This ERP response has been associated to the P3b response observed in scalp EEG in response to global novelty, as it is observed in regions identified as sources of the P3b (Bekinschtein et al., 2009). It has also been reported in previous intracranial studies (Smith et al., 1990; Baudena et al., 1995; Halgren et al., 1998). Moreover, responses to global novelty also included ii) a sustained decrease in beta (13-25 Hz) power starting on average 258 ms after the onset of the critical sound, and iii) a prolonged increase in PPC in the alpha band (8-12 Hz) beginning from 160 ms after the onset of the fifth sound, particularly in pairs of frontal electrodes. A similar effect was observed with the induced PPC, which suggests genuine changes in functional connectivity in this time window, over and above the mere propagation of stimulus-induced activation. These results need to be confirmed in future studies, as the high dependency between pairs of electrodes did not allow us to correct for multiple comparison across pairs of electrodes, but only across time and frequency samples. Moreover, as for all intracranial EEG studies, estimation of network connectivity was limited by the localization of recording sites in each patient. Notably, it would have been interesting to refine our understanding of the interactions between temporal and frontal electrodes, but only three patients presented recording sites in both of these regions.

6.2. Local processing versus long-range interactions

Our time-frequency results revealed opposite patterns of spectral power for local and global novelties. Local novelty elicited an early and transient increase in high gamma (60-120 Hz) power within the auditory cortex, in agreement with a previous intracranial study using a classical oddball paradigm (Edwards et al., 2005). Conversely, a distributed, late and sustained decrease in power in the beta (13-25 Hz) band was observed in response to global novelty.

Interestingly, high gamma activity has been proposed as a signature of local spiking activity (Pesaran et al., 2002; Nir et al., 2007; Ray and Maunsell, 2010), whereas modulations of power in the beta band are thought to reflect long-range interactions (Kopell et al., 2000). Moreover, a recent framework (Donner and Siegel, 2011) associates high gamma activity with encoding functions and lower frequency (including beta band) oscillations with integrative functions. Our results are in agreement with this framework, as the global effect involves the integration of sounds over several trials to identify the relevant rule, whereas the local effect reflects the encoding of changes in basic features of the sound. Note, however, that the PPC analysis did not allow us to identify long-range interactions in response to global novelty. This may be related to the few long-range pairs of electrodes present in this dataset, as for example only three patients were implanted both in the temporal and the frontal lobes.

6.3. Partition of the MMN into two successive events

Our observation of two successive events in our ERP analysis of responses to local novelty supports the partition of the MMN in two main components: 1) an early automatic, non-conscious, and short-lived echoic memory system located within the auditory cortex, termed 'early MMN' (Pegado et al., 2010) and 2) a later response extending beyond early auditory areas, and showing a stronger resistance to increases in the temporal spacing of sounds, corresponding to the N2b component or 'late MMN' (Pegado et al., 2010). These results are in agreement with previous intracranial studies using a classical oddball paradigm (Halgren et al., 1995a). Interestingly, the N2b has been associated with attention (Näätänen and Gaillard, 1983), contrary to the 'early MMN' which is thought to reflect automatic processing of novelty. In Pegado et al.'s study (2010), the authors show that when increasing the SOA between two successive sounds beyond 1000ms in a traditional odd-ball paradigm, the disappearance of the 'early MMN' actually corresponds to the occurrence of a similar novelty response present both for deviant and standard trials. By contrast, the 'late MMN' latency was delayed, but this component did not disappear. Pegado et al. proposed a model accounting for these results: the early MMN would reflect accumulation of evidence based on echoic memory representations,

whereas the late MMN would be related to accumulation of evidence on longer time scales. In our study, the early MMN would thus be absent in the global novelty condition because the relevant time scales are longer than in the local novelty condition. Conversely, given the resistance of the 'late MMN' to SOA, it would not be surprising to observe its presence in response to global novelty.

In the present study, no MMN was recorded in frontal electrodes, in agreement with some previous intracranial studies (Baudena et al., 1995; Edwards et al., 2005). Nevertheless, because of the non-uniform sampling of our recording sites over the frontal lobe, it cannot be excluded that a frontal generator of the MMN exists but was missed, as suggested by some scalp EEG studies (Giard et al., 1990; Rinne et al., 2000; Opitz et al., 2002). Note that in a study including only two implanted patients (Bekinschtein et al., 2009), using less conservative statistical thresholds, we previously reported 3 frontal electrodes with a local effect, in agreement with other intracranial studies (Liasis et al., 2001; Rosburg et al., 2005).

6.4. Global effect and conscious access

The paradigm used in this study was designed to dissociate two different processes, underlying local and global novelty detection. In a previous experiment using this paradigm, we showed that while the response to local novelty was still observed under conditions of inattention (engagement in a concurrent difficult RSVP task), the response to global novelty disappeared and subjects could not report the existence of violations of the global regularity in EEG and MEG studies (Bekinschtein et al., 2009). Moreover, this paradigm was used in patients suffering from disorders of consciousness (DOC) and the response to global novelty was only observed in patients showing signs of consciousness (Bekinschtein et al., 2009; Faugeras et al., 2011, 2012; King et al., 2013). The P3b component has also been proposed as a marker of conscious access in other paradigms (Sergent et al., 2005). Note, however, that in the present study, patients were instructed to count global deviant trials. The responses to global novelty could thus be reflect

downstream processes relative to conscious access per se (Aru et al., 2012; Sergent and Naccache, 2012).

To distinguish between these "downstream" processes and signatures of conscious access, it is interesting to compare our results with studies using a different task. First, Wacongne et al (2011) used a passive version of the Local/Global paradigm and found similar ERP signatures of local and global novelty. Moreover, in a completely different study contrasting masked and unmasked words, Gaillard et al. (2009) explored ERPs, time-frequency changes and connectivity patterns in intracranial EEG, - similarly to the approach taken in the present study -, and reported closely related findings. Conscious processing of unmasked words was associated with sustained event-related components, notably in the frontal cortex, a decrease in power in the beta band and a late increase in connectivity in the alpha and beta band. The single discrepancy with our findings is the small proportion of electrodes in which we observed the three signatures of the late global effect. This could be explained by the fact that the global effect corresponds to a more subtle contrast than the masked/unmasked contrast. Indeed, whereas the masked/unmasked contrast compared a 'no stimulus' condition with a 'stimulus present' condition, we contrasted here two conditions associated with conscious perception: global deviant stimuli being perceived as violations of the global regularity, and global standard stimuli as regular trials.

Nevertheless, we conclude by highlighting the striking similarity between our results and the three signatures of conscious perception identified by Gaillard et al, despite the profound difference in sensory modalities (audition versus vision), in stimuli (words versus tones) and tasks. This may point to a general common mechanism of conscious access (Sergent and Naccache, 2012).

7. Acknowledgments

Conflict of interest: The authors declare no competing financial interests.

Funding: This work was supported by the program 'Investissements d'avenir' ANR-10-IAIHU-06, an AXA Research Fund grant to IEK, a Direction Générale de l'Armement (DGA) grant to JRK, a Ministère de l'Enseignement Supérieur et de la Recherche grant to FM, a Rubicon grant from the Netherlands Organization for Scientific Research (NWO) to SvG, an 'Equipe FRM 2010' grant of Fondation pour la Recherche Médicale (FRM) to LN and an European Research Council (ERC) senior grant "NeuroConsc" to SD.

Acknowledgments: We thank Séverine Samson and Isabelle Jegourel for their help in gathering clinical information and all the patients for their time and kindness.

CHAPTER 3. DO ACQUISITION AND TRANSFER OF STRATEGY REQUIRE CONSCIOUS ACCESS?

1. Presentation of the article

In this second study, we pursue our work on how subjects deal with inconsistencies in their environment. We designed an experimental paradigm derived from the Stroop task, in which 80% of trials were incongruent. In this condition of high conflict frequency, behavioral measures and EEG responses were collected to assess adaptation to conflict. Moreover, conscious and non-conscious trials were compared to test the existence of conflict adaptation in the absence of awareness. This study aimed at understanding how subjects adapt their behavior after having detected regularities in their environment and whether this adaptation can be triggered by unconscious stimuli.

2. Abstract

It has been assumed for long that cognitive control processes require consciousness, but empirical evidence is contradictory. In the present study, we investigated strategic adaptation to conflict in masked and unmasked trials, using behavioral measures and event-related potentials (ERPs). We studied not only the acquisition of strategies in masked trials, but also the transfer of a strategy developed in unmasked trials to masked trials, and the trial-to-trial dynamics of strategic processing in unmasked trials. To do so, we used an experimental paradigm derived from the Stroop task, with masked and unmasked baseline (50% incongruent trials) and mostly-incongruent (80% incongruent trials) blocks. Importantly, we included masked trials in unmasked blocks, in order to study the ability to transfer strategies. In unmasked trials, we found clear evidence of strategic adaptation to conflict, both in reaction times and in ERPs (decreased N2 and increased P300). In masked trials, on the other hand, we found no evidence of behavioral adaptation to conflict, but a modulation of the P300 was present in masked trials included in unmasked blocks, suggesting the existence of a transfer of strategy. Finally, trial-to-trial analyses in unmasked trials revealed an interesting pattern suggestive of dynamic subjective adherence to the instructed strategy.

Key words: cognitive control, consciousness, Stroop, blockwise adaptation

3. Introduction

Cognitive control refers to our ability to rapidly and flexibly adapt our behavior depending on our current goals and environment. Several paradigms have been designed to study this process, including task switching paradigms in which subjects have to respond to the same stimuli in different ways according to the context (Monsell, 2003), or stop-signal tasks in which a stimulus requires the participant to stop his action (Logan, 1982). Cognitive control can also be observed in interference tasks. A classic example of such a task is the Stroop task (Stroop, 1935). In this paradigm, subjects are presented with color words (e.g. red) printed in a colored ink and they have to respond to the ink color, independently of the word meaning. The ink color and the color associated with the meaning of the word can either be congruent ("red" printed in red) or incongruent ("red" printed in green). In incongruent trials, there is a response conflict between the relevant information (the color of the ink) and the irrelevant information (the meaning of the word), and participants are thus typically slower to respond than in congruent trials. This effect is known as the Stroop effect. Interestingly, it can be modulated depending on the context: the influence of irrelevant information is reduced when response conflict is frequent (e.g. in a block of trial - Logan and Zbrodoff 1979) or recent (e.g. in the previous trial - Gratton et al. 1992). This adaptation to response conflict is a typical example of cognitive control.

A debated issue is whether cognitive control can be triggered unconsciously (for reviews see Desender and Van den Bussche 2012; Kunde et al. 2012; van Gaal et al. 2012; Ansorge et al. 2014). Indeed, cognitive control has been proposed to be only reserved for consciousness (Dehaene and Naccache, 2001; Jack and Shallice, 2001). But recent experimental evidence challenges this view. For example, van Gaal et al. (2009) showed that participants slowed down their responses following an unconscious stop signal, suggesting that it was partially processed but not enough to fully inhibit their response. However, experimental evidence for unconscious cognitive control is ambiguous: triggering cognitive control by explicit events such as a stop signal or a task-switching signal (Lau and Passingham, 2007) seems to be possible even if this

signal is unconscious, but when the event triggering cognitive control is derived from regularities in the environment, such as the frequency of response conflict, the results are much more disputed (for review see Kunde et al. 2012). Note that explicit events such as stop signals have to be learned consciously, before they can be used in unconscious trials. Conflict adaptation on a trial-to-trial basis when the conflict itself remains unconscious have been studied for example by Kunde (2003). In this study using meta-contrast masking, adaptation to conflict was observed only after a conscious trial, which suggests that an unconscious conflict cannot alter the processing of the next trial. However, van Gaal et al. (2010) used a similar paradigm, shortening the inter-trial interval and omitting a warning sound at the beginning of each trial. They observed conflict adaptation on the current trial, independently of the visibility of the previous trial. This study thus suggests that conflict adaptation could be triggered by an unconscious conflict (see also Bodner and Mulji 2010; Desender et al. 2013), but it can also be interpreted alternatively as the consequence of the awareness of a meta-cognitive information, such as reaction time slowing (Ansorge et al., 2011). Another interesting example is the work of Merikle and colleagues. They showed in a series of studies that subjects used predictive strategies based on blockwise conflict frequency only when the conflict was conscious (Merikle and Cheesman, 1987; Merikle et al., 1995; Merikle and Joordens, 1997). These studies were based on a variant of the Stroop task: subjects were presented with a color prime word (red or green in grey color) for 33 ms that was followed by a target which consisted of seven colored ampersands, and they were asked to respond to the color of this target as quickly as possible. A pattern mask consisting of seven ampersands in gray color was presented between the prime and the target, either immediately after the prime in the unconscious condition or after a delay of 134 ms in the conscious condition. The stimulus onset asynchrony (SOA) between the prime and the target was held constant across conscious and unconscious trials (300 ms). The prime and the target were incongruent in 75% of the trials. The authors found an interaction between the visibility of the prime and the prime-target congruency: subjects responded faster on incongruent than on congruent trials in the conscious condition (reverse Stroop effect), whereas

they responded slower on incongruent than on congruent trials in the unconscious condition (Stroop effect). These studies suggest that strategic blockwise adaptation to conflict can only be observed for conscious stimuli and were later replicated (Daza et al., 2002). However, these results have been challenged by several studies showing blockwise adaptation to unconscious conflict frequency manipulations (Bodner and Masson, 2001; Jáskowski et al., 2003; Klapp, 2007; Bodner and Mulji, 2010).

In addition to behavioral measures, it is possible to study conflict processes using various measures of neural activity (Sternberg, 2001), and in particular in relation to frontal lobe activity. Indeed, the activity in the anterior cingulate cortex (ACC) has been associated with response conflict monitoring using functional MRI (fMRI) and Positron Emission Tomography (PET) (Pardo et al., 1990; Carter et al., 2000; Botvinick et al., 2001, 2004; Ridderinkhof et al., 2004a). The dorsolateral prefrontal cortex (DLPFC) and the supplementary motor area (SMA) have also been associated with conflict adaptation (Kerns et al., 2004; Ridderinkhof et al., 2004a; Bonini et al., 2014). In addition to these studies providing a good spatial resolution, electroencephalography (EEG) was used to better understand the time course of conflict adaptation. Two event-related potentials (ERPs) were identified. The first one is the fronto-central N2 component peaking between 200 ms and 350 ms after stimulus onset, which, according to source reconstruction studies, is generated in the ACC (van Veen and Carter, 2002; Nieuwenhuis et al., 2003; Folstein and Van Petten, 2008). In agreement with the role of ACC in the conflict monitoring theory (Botvinick et al., 2001), the amplitude of this component is higher on incongruent compared to congruent trials (Danielmeier et al., 2009; Forster et al., 2010; Clayson and Larson, 2011). The second component modulated by congruence is the centro-parietal P300, peaking 300 to 500 ms after stimulus onset (Polich, 2007; Clayson and Larson, 2011; Frühholz et al., 2011). Its functional significance is less clear than the N2 components. It has been associated to response inhibition (Frühholz et al., 2011) or response evaluation (Kopp et al., 1996).

In the present study, using both behavioral and EEG data, we aimed at exploring in more details the blockwise adaptation to conflict and its relation to consciousness, with an experimental paradigm similar to the one used Merikle and colleagues' studies (Merikle and Cheesman, 1987; Merikle et al., 1995; Merikle and Joordens, 1997). Importantly, we introduced two baseline conditions (masked and unmasked) in order to detect any change in the Stroop effect relative to the high proportion of incongruent trials. Indeed in previous studies, only a reversal of the Stroop effect was considered as reflecting strategic processing (Merikle et al., 1995; Merikle and Joordens, 1997; Daza et al., 2002; Ortells et al., 2003), whereas the comparison with a condition with 50% incongruent could reveal more subtle effects. Moreover, we explicitly instructed subjects about the high proportion of incongruent trials (Daza et al., 2002), in order to test the mere application of the strategy that subjects did not have to infer. We also decided to use a short stimulus-onset asynchrony (SOA) between the prime and the target, compared to previous studies in which no strategic effect was found in masked trials (Merikle et al., 1995; Merikle and Joordens, 1997; Daza et al., 2002; Ortells et al., 2003), as subliminal priming is known for decreasing with time (Greenwald et al., 1996; Dupoux et al., 2008). Note also that in the studies showing blockwise adaptation to conflict in masked trials, a short SOA between the prime and the target was used (Bodner and Masson, 2001, 2004; Jáskowski et al., 2003; Klapp, 2007). Finally, we added a new experimental condition for masked trials by comparing blocks exclusively composed of masked trials with blocks including a minority of masked trials randomly intermixed with unmasked trials. We speculated that even if no strategic processing was observed in fully masked blocks, - as reported by Merikle and colleagues' studies (Merikle and Cheesman, 1987; Merikle et al., 1995; Merikle and Joordens, 1997) -, the strategy deployed on unmasked trials may well transfer to masked trials (Naccache, 2006). Indeed several recent studies demonstrated that unconscious processing may be modulated by top-down influence such as endogenous spatial and temporal attention (Kentridge, 1999; Naccache et al., 2002).

Using this adapted paradigm, we first aimed at replicating the strategic blockwise adaptation to conflict in unmasked trials and analyzed the trial-to-trial dynamics of this strategy. Second, we

studied blockwise adaptation to conflict in masked trials with explicit instructions and tested the mere application of an instruction in masked blocks. Finally, we tested the hypothesis that once established on unmasked trials, the strategic use of the prime could be transferred to masked trials, by including 30% masked trials in unmasked mostly-incongruent blocks.

4. Materials and methods

4.1. Subjects

Twenty-one right-handed subjects (10 women, age mean = 23.2 years, sd = 2.4) were included in this study. They all reported normal or corrected-to-normal visual acuity. The experiment was approved by the ethical committee of the Pitié-Salpétrière hospital. All subjects gave their written informed consent and were paid 40€ to participate in the experiment.

4.2. Stimuli and Procedure

Stimuli were presented against a gray background at the center of a 17-inch LCD Dell screen (frequency 60 Hz, resolution 1290 x 1024). The procedure is illustrated in Fig. 3.1. We used a modified version of the Stroop task. Each trial started by a fixation cross that lasted 1s, then a pre-mask composed of the characters &$£@#%$ ordered randomly was presented for 200 ms. A cue consisting of a color word (green or blue) was then presented for 33 ms and was followed by either a post-mask composed of the same characters as the pre-mask but in another random order in masked trials, or by a blank screen in unmasked trials, during 100 ms (stimulus-onset asynchrony (SOA) = 133 ms). Finally, a target consisting in a blue or green string of ampersands (&&&&&&) was presented and subjects had to respond to the color of this string by pressing a button with their left or right hand. The color/hand correspondence was counterbalanced across subjects.

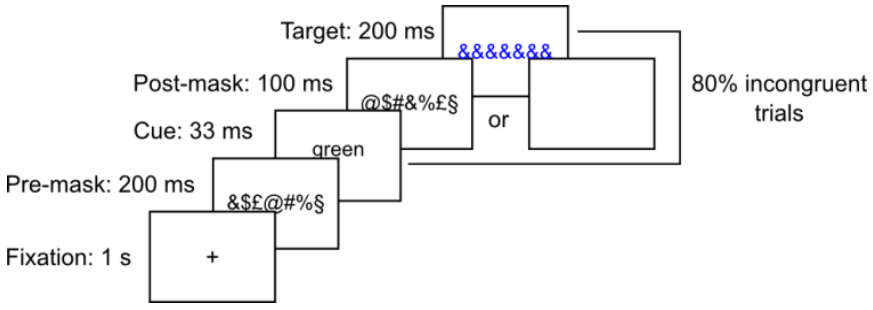

FIGURE 3.1: EXPERIMENTAL PARADIGM

We used a modified version of the Stroop task. Subjects were first presented with a color word as a cue (blue or green) and they had to respond to the color of a meaningless string. The cue could either be masked or unmasked. There were 4 types of blocks: masked baseline (50% incongruent trials), unmasked baseline (50% incongruent trials), mostly-incongruent masked blocks (80% incongruent trials) and unmasked mostly-incongruent blocks (80% incongruent trials) which included 70% unmasked trials and 30% masked trials.

The cue and the target were either congruent (the word corresponded to the color of the target) or incongruent (the word did not correspond to the color of the target). Subjects were presented with 10 critical blocks of 100 trials containing a high proportion of incongruent trials (80% - mostly-incongruent condition). They were told about this high proportion of incongruent trials and were instructed to use the cue word to prepare the response corresponding to the opposite color, as this was the best strategy to answer rapidly and accurately in most trials. Among these 10 blocks, 2 blocks contained only masked trials and the 8 others contained 70% of unmasked trials and 30% of masked trials. These blocks allowed us to test the development of the strategy on fully masked blocks but also on masked trials embedded in unmasked blocks, in which the strategy established on unmasked trials could be transferred to masked ones. Finally, subjects were presented with 2 blocks of 100 unmasked trials with 50% incongruent trials (unmasked baseline condition) and 2 blocks of 100 masked trials with 50% incongruent trials (masked baseline condition). These blocks of baseline were presented at the beginning (one block of each) and at the end of the experiment (one block of each). The order of the unmasked and masked baseline was random. Feedback on reaction time and accuracy was provided to subjects at the end of each block. In total, subjects performed 1400 trials.

At the very end of the experiment, subjects performed a discrimination task on the cue, in 2 distinct blocks of 100 trials: the first one comprised only masked trials and the second one comprised 70% of unmasked trials and 30% of masked trials, as in the main task. In these blocks, 50% of trials were congruent. Their order of presentation was counterbalanced across subjects. In total, the experiment lasted about 45 min.

4.3. Behavioral data analysis

Incorrect trials and correct trials with reaction times faster than 200 ms or slower than 800 ms (representing 2.3% of trials) were excluded from all analyses on reaction times. Behavioral data were analyzed with MATLAB R2011a (The MathWorks, Inc.) and the Statistical toolbox to perform t-tests and repeated-measures ANOVA when comparing the different conditions.

Four subjects were excluded as they showed a negative Stroop effect in the unmasked baseline condition and an additional subject was excluded due to excessive artefacts on the EEG signal. So the results presented are based on 16 subjects.

4.4. EEG recordings and analyses

EEG activity was continuously recorded using a 256-channel EEG net (Electrical Geodesics Inc, Eugene, OR). The signal was sampled at 250 Hz.

All analyses were conducted using Fieldtrip toolbox (Oostenveld et al., 2011). EEG data were filtered between 0.5 and 30 Hz and then epoched from -400 ms to 900 ms relative to the onset of the target. Then, the data were visually inspected for artefacts not related to blinks. Bad channels were interpolated. Independent components analysis was computed and components containing blinks or oculomotor artefacts clearly dissociable from brain activity were subtracted from the data. Finally, the data were re-referenced to common average and baseline corrected over 200 ms before the cue, between -400 and -200 ms relative to the onset of the target.

To improve signal-to-noise ratio and based on previous literature (Folstein and Van Petten, 2008; Hanslmayr et al., 2008; Frühholz et al., 2011; Jiang et al., 2013) about the components of interest (N2 and P3), a spatial region-of-interest (ROI) including 15 channels was created around Cz.

To assess the significance of differences between congruent and incongruent trials in the different conditions, we used dependent sample t-tests across subjects at each time point in the ROI around Cz. Then correction for multiple comparisons over time was performed, by nonparametric cluster-based method (Maris and Oostenveld, 2007) . We computed N = 1000 permutations by shuffling trial labels. Then, for each permutation, dependent sample t-tests were performed at each time sample. All samples with a t-value corresponding to a p-value smaller than 0.05 were clustered in connected sets on the basis of adjacency in time. Then the cluster statistic was computed by taking the sum of the t-values within each cluster. The cluster-corrected threshold was obtained by computing the permutation distribution of the maximum

cluster statistic and taking the c+1 largest member of this distribution, with c is equal to αN rounded down. In the original data, only clusters with a cluster statistic higher than this threshold were identified as significant. All the reported p-values are corrected for multiple comparisons.

5. Results

5.1. Behavior

Adaptation to conflict frequency was studied in 5 conditions: unmasked and masked baseline conditions (with 50% incongruent trials), unmasked and masked mostly-incongruent conditions (with 80% incongruent trials), as well as masked trials included in unmasked mostly-incongruent blocks.

The unmasked baseline condition was used to assess the existence of the Stroop effect, as measured by the difference in reaction times between incongruent and congruent trials, and four subjects were excluded because they showed a negative Stroop effect in this condition (mean Stroop effect with these subjects = 23 ms; mean Stroop effect without these subjects = 31 ms). Then the masked baseline condition was analyzed to test that the masked prime was correctly processed. There was a significant Stroop effect in the masked baseline condition (50% incongruent trials) (mean = 5.5 ms; $t(15) = 2.52$, $p = 0.02$). Moreover, the visibility of the masked prime was analyzed through d' scores on the discrimination task and subjective reports at the end of the experiment, according to which participants were not able to identify the cues. In the block containing only masked prime, d' scores significantly different from 0 (mean = 0.25, sd = 0.31; t-test, $t(15) = 3.16$, $p = 0.006$), but they were not correlated with the Stroop effect in the masked baseline condition (Pearson correlation, $r = -0.11$, $p = 0.66$) and the intercept was significantly different from 0 (estimate = 0.27; $t(15) = 2.81$, $p = 0.013$), suggesting a genuine interpolated unconscious effect under null d' (Greenwald et al., 1995). In the block containing only 30% of masked trials, d' scores were significantly different from 0 (mean = 1.13, sd = 0.6; t-test, $t(15) = 7.54$, $p = 1.7*10^{-6}$), but they were not correlated ($r = -0.02$, $p = 0.93$) with the Stroop effect for the masked trials included in unmasked blocks (mean Stroop effect = 6.6 ms; $t(15) = 1.5$, $p = 0.14$) and the intercept was significantly different from 0 (estimate = 1.14; $t(15) = 6.8$, $p = 8.5*10^{-6}$). These results suggest that the cue was processed and affected behavior, even in the masked condition.

We performed a repeated-measure ANOVA on the Stroop effect with visibility (masked/unmasked) and incongruent proportion (baseline/mostly incongruent) as within-subjects factors and subjects as a random factor (Fig. 3.2a). Note that the masked trials included in the unmasked mostly-incongruent blocks could not be included in this analysis. We found a main effect of visibility ($F(1,60) = 18.9$, $p = 6*10^{-4}$) as well as a significant interaction between visibility and incongruent proportion ($F(1,60) = 12.21$, $p = 0.003$). The main effect of incongruent proportion showed a trend ($F(1,60) = 3.55$, $p = 0.079$). The same analysis on error rates showed no significant effect (all $F<1$), so we decided to focus on reaction time data. Restricted analyses showed that in the unmasked mostly-incongruent condition, there was a significant decrease in the Stroop effect relative the unmasked baseline condition ($t(15) = 2.93$, $p = 0.01$ – Fig. 3.2a). This effect suggests that participants used the prime strategically in the unmasked mostly-incongruent blocks: they tended to be faster to respond to incongruent trials

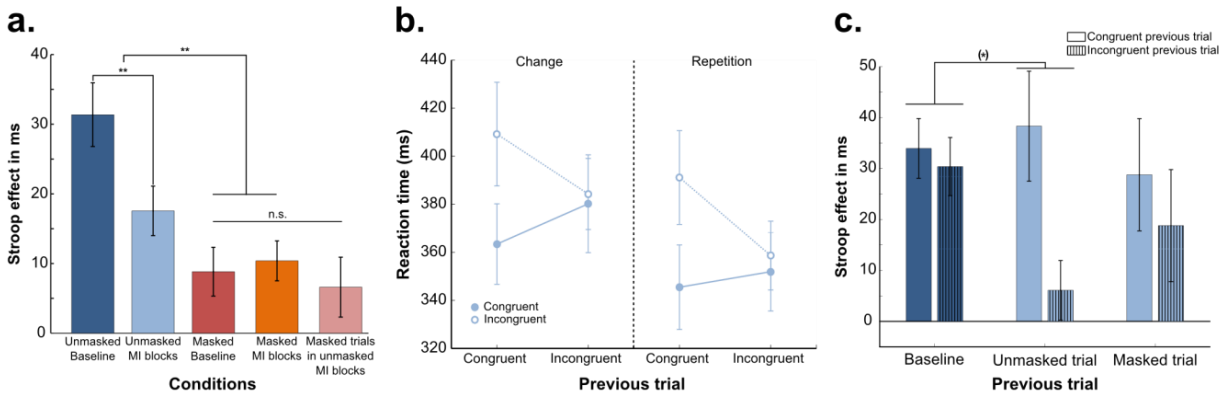

FIGURE 3.2: BEHAVIORAL RESULTS

(a) Stroop effect, as measured by the difference in reaction times between incongruent and congruent trials, is presented in the different experimental conditions. The Stroop effect is strongly reduced in the unmasked mostly-incongruent (MI) blocks compared to baseline, but no difference relative to baseline is observed for masked trials.

(b) Reaction time is presented according the current trial congruence and the previous trial congruence, both for physical repetition and change across the previous and current trials. A significant interaction between current trial congruence and previous trial congruence is observed even when excluding physical repetition ($F(1,60) = 4.6$, $p = 0.04$), suggesting that these results reflect trial-to-trial adaptation to conflict.

(c) Stroop effect for unmasked trials is presented according to the visibility and the congruence of the previous trial. It is significantly different from baseline only when the preceding trial is unmasked and incongruent, suggesting an online evaluation of the instructed strategy.

than to congruent trials in this condition than when they processed the prime automatically in the baseline. On the contrary, no significant change relative to baseline was observed in the masked conditions, neither in fully masked blocks, nor in masked trials included in the unmasked blocks (one-way ANOVA, $F(2,45) = 0.46$, $p = 0.63$ – Fig. 3.2a). These results highlight the difference between the masked and unmasked conditions: the prime seems to be strategically used only when it is unmasked.

We thus decided to focus on unmasked trials in order to better understand the fine dynamics of this adaptation to conflict. First, we analyzed the influence of the previous trial congruence on the Stroop effect in unmasked trials preceded by an unmasked trial. We performed a repeated-measure ANOVA on reaction times in these trials with current trial congruence and previous trial congruence as within-subject factors and subjects as a random factor. We found a main effect of current trial congruence ($F(1,60) = 34.23$, $p = 3.19*10^{-5}$) and a main effect of previous trial congruence ($F(1,60) = 11.2$, $p = 0.004$) and the interaction between current trial congruence and previous trial congruence showed a trend ($F(1,60) = 4.25$, $p = 0.057$). This result shows the existence of trial-to-trial adaptation (Gratton et al., 1992). Interestingly, restricted analyses showed that incongruent trials were particularly affected by the previous trial: responses to incongruent trials preceded by an incongruent trial were faster than when preceded by a congruent trial (371 ms vs. 399 ms; t-test, $t(15) = -3.17$, $p = 0.006$), but responses to congruent unmasked trials were not affected by the previous trial congruence (360 ms when the previous trial was congruent vs. 365 ms when the previous trial was incongruent; t-test, $t(15) = -0.56$, $p = 0.57$). Moreover, following a congruent unmasked trial, the Stroop effect was not significantly different from the unmasked baseline ($t(15) = -0.56$, $p = 0.58$), whereas following a incongruent unmasked trial, the Stroop effect was significantly smaller than in the unmasked baseline condition ($t(15) = 4.08$, $p = 9*10^{-4}$). To exclude the role of stimulus repetition in this trial-to-trial conflict adaptation (Mayr et al., 2003), we performed a repeated-measure ANOVA including current trial congruence, previous trial congruence and physical repetition as within-subject factors and subjects as a random factor (Fig. 3.2b). We found a main effect of

current trial congruence ($F(1,120) = 39.29$, $p = 1.5*10^{-5}$), a main effect of previous trial congruence ($F(1,120) = 7.13$, $p = 0.01$), a main effect of physical repetition ($F(1,120) = 6.41$, $p = 0.02$) and the interaction between current trial congruence and previous trial congruence ($F(1,120) = 6.01$, $p = 0.02$). Importantly, the triple interaction between current trial congruence, previous trial congruence and physical repetition was not significant ($F(1,120) = 0.03$, $p = 0.87$), indicating that the trial-to-trial adaptation was not different for physical repetitions and changes. We performed restricted analyses on physical repetitions and changes separately with two repeated-measure ANOVAs on reaction times with the current trial congruence and the previous trial congruence as within-subject factors and subjects as a random factor. We observed a main effect of the current trial congruence for both analyses ($F(1,60) = 17.8$, $p = 7*10^{-4}$ and $F(1,60) = 24.9$, $p = 2*10^{-4}$ respectively), and the main effect of the previous trial congruence was significant only for repetitions ($F(1,60) = 5.6$, $p = 0.03$). Interestingly, the critical interaction between the current trial congruence and the previous trial congruence was significant both for changes and repetitions ($F(1,60) = 4.6$, $p = 0.04$ and $F(1,60) = 4.9$, $p = 0.04$ – Fig. 3.2b). Thus trial-to-trial adaptation to conflict does not depend on physical repetitions and really reflect cognitive control processes. Taken together, all these results suggest that following an incongruent trial (i.e. a trial in agreement with the instructed strategy), subjects tend to show a strongly reduced Stroop effect, whereas their Stroop effect is similar to the one observed in the baseline condition following a congruent trial (i.e. a trial that contradicts the instructed strategy). It suggests that subjects are evaluating the validity of the instructed strategy on each trial and adapt their behavior accordingly.

It is thus interesting to see how masked trials interfere with the application of the strategy. We analyzed the response in unmasked trials as a function of the visibility of the previous trial. When compared to the unmasked baseline condition, the Stroop effect for unmasked trials preceded by unmasked trials was strongly reduced ($t(15) = 3.88$, $p = 0.001$), but was similar to baseline for unmasked trials preceded by masked trials ($t(15) = 0.67$, $p = 0.51$). Moreover, for unmasked trials preceded by an unmasked trial, a repeated-measure ANOVA on the Stroop effect

with previous trial congruence and incongruent proportion as within-subject factors and subjects as a random factor showed a main effect of incongruent proportion ($F(1,60) = 5.22$, $p = 0.03$) and critically the interaction between the previous trial congruence and incongruent proportion almost reach significance ($F(1,60) = 4.02$, $p = 0.06$ – Fig. 3.2c). For unmasked trials preceded by a masked trial, none of these effects reached significance (both $F < 1.1$ – Fig. 3.2c). These results suggest that following a masked trial in the blocks containing mostly unmasked trials, independently of its congruence, subjects behave as after an unmasked congruent trial and do not use the prime strategically.

In summary, we found a significant effect in the masked and unmasked baseline conditions, indicating that the prime was properly processed in both cases. Strategic blockwise adaptation was only found for unmasked trials and no evidence of acquisition or transfer of strategy could be shown for masked trials. Interestingly, visibility and congruence of the previous trial strongly influenced the use of the strategy, suggesting that subjects evaluated the validity of the strategy from trial to trial.

5.2. Event-related potentials

EEG activity around Cz in the different conditions is displayed in Fig. 3.3. Two main event-related potential (ERP) components modulated by congruence were identified. An N2 component significantly modulated by congruence ($p_{corr} = 0.003$) was observed in the unmasked baseline condition from 256 ms to 368 ms after the target onset (peak difference = -0.98, peak latency = 324 ms – Fig. 3.3a). A P300 component significantly modulated by congruence ($p_{corr} = 0.003$) was observed from 408 ms to 568 ms after the target onset in the mostly-incongruent unmasked condition (peak difference = 0.57, peak latency = 476 ms – Fig. 3.3b). We performed a repeated-measure ANOVA on the amplitude of these components with component (N2/P300) and condition (baseline/mostly-incongruent) as within-subject factors and subjects as a random factor. Interestingly, the interaction between these factors was significant ($F(1,60) = 7.55$, $p = 0.01$ – Fig. 3.3c). For masked trials, no significant difference between congruent and incongruent trials was observed in the baseline (Fig. 3.3d) and in the masked mostly-incongruent blocks, the

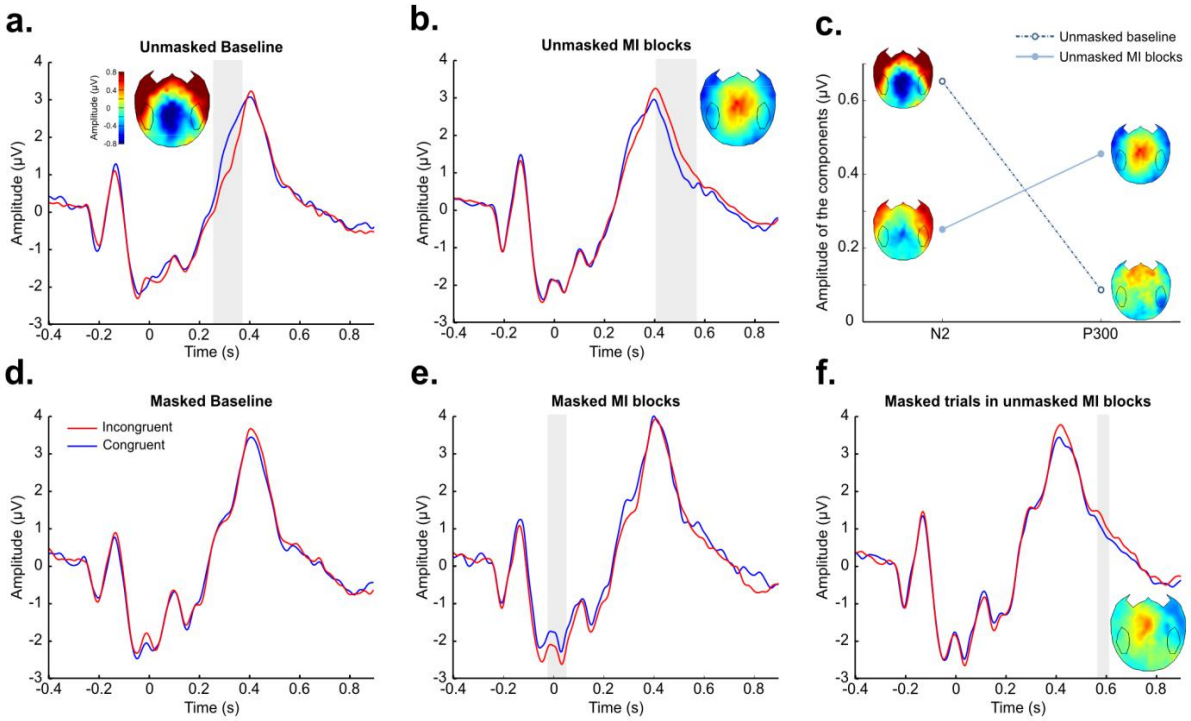

FIGURE 3.3: TWO ERP COMPONENTS: N2 AND P300

In panels (a), (b), (d), (e) and (f), ERP analyses from a cluster of 15 electrodes around Cz in different experimental conditions are presented, with incongruent trials in red and congruent trials in blue. The gray shadings indicate time periods of statistical significance (pcorr < 0.05) and scalp topographies of the difference between incongruent and congruent trials show the effects of congruence during these time periods.

(a) In the unmasked baseline, the N2 component is modulated by congruence from 256 ms to 368 ms after the target onset.

(b) In unmasked mostly-incongruent (MI) blocks, the P300 component is modulated by congruence from 408 ms to 568 ms after the target onset.

(c) The amplitude of the N2 and P300 components was computed as the absolute value of the difference between congruent and incongruent trials in the time windows of interest for each component. We found a significant interaction ($F(1,60) = 7.55$, $p = 0.01$) between the amplitude of the components and the experimental condition (unmasked baseline – dotted line/unmasked mostly-incongruent blocks – continuous line). Scalp topographies of the difference between incongruent and congruent trials for the time window of interest for each component and in each condition are plotted next to the corresponding point in the plot.

(d) No significant difference was observed in the EEG between congruent and incongruent trials in the masked baseline condition.

(e) In masked mostly-incongruent blocks, the only significant difference was outside the a priori time windows of interest (from -24 ms to 48 ms after the target onset).

(f) Interestingly, in the masked trials included in the unmasked mostly-incongruent blocks, a modulation of the P300 by congruence was observed from 564 ms to 608 ms after the target onset, similar to the signature observed for unmasked trials in mostly-incongruent blocks.

only difference reaching significance (p_{corr} = 0.02) was from -24 ms to 48 ms, outside the a priori time windows of interest (Fig. 3.3e). But interestingly, a P300 significantly modulated by congruence (p_{corr} = 0.04) was observed for masked trials in mostly-incongruent unmasked blocks from 564 ms to 608 ms (Fig. 3.3f). This modulation suggests the existence of a transfer of the strategy in these trials, even if this effect could not be observed in the behavior.

Finally, for unmasked trials in the mostly-incongruent condition, we analyzed the ERP around Cz taking into account the congruence and visibility of the previous trial. Interestingly, we found a significant modulation of the P300 component by the congruence of the current trial only when the previous trial was unmasked and incongruent (i.e. when the previous trial supported the instructed strategy), both when the current trial was unmasked ($t(15) = 4.63$, $p = 0.0003$) and when it was masked ($t(15) = 2.9$, $p = 0.01$).

6. Discussion

In the present study, we analyzed the interplay between cognitive control and consciousness, by exploring the existence of strategic adaptation to conflict for masked and unmasked trials in mostly incongruent blocks, using an experimental paradigm adapted from the Stroop task. We observed blockwise adaptation to conflict in unmasked trials. However, despite clear behavioral evidence of unconscious processing in masked trials, blockwise adaptation to conflict was not observed in these trials. This negative result was found both when masked trials were included in fully masked blocks and when they were embedded within mostly unmasked blocks. Thus, no evidence for acquisition or transfer of the strategic use of the prime was found in the behavior. Nevertheless, in the ERP analysis, the modulation of the P300 component by prime-target congruence observed in 'mostly incongruent' blocks for unmasked trials was also present for masked trials embedded in unmasked blocks. This last finding suggests the existence of a transfer of strategy from unmasked trials to masked trials. Finally, detailed analyses of trial-to-trial adaptation to conflict in unmasked trials revealed original effects related to the visibility and the congruence of the previous trial.

In unmasked trials, as Merikle and colleagues (Merikle et al., 1995; Merikle and Joordens, 1997), we found a significant impact of the strategy on the Stroop effect in the mostly incongruent condition. Note however that in our case the modulation of the Stroop effect compared to the baseline did not extend to a full inversion as reported by Merikle and colleagues. This difference can be explained by the short SOA we used (133 ms), in order to keep masked and unmasked trials as similar as possible and to maximize unconscious processing. However, it has been shown that the SOA between the prime and the target should be at least 300 ms to observe a reversal of congruence effect (Merikle et al., 1995; Merikle and Joordens, 1997; Daza et al., 2002). Nevertheless, including a baseline condition allowed us to show the existence of strategic processing in unmasked trials, even with a short SOA.

Moreover, the ERP analysis in unmasked trials revealed that two components were modulated by the task. Indeed, we observed that the N2 was suppressed in mostly incongruent blocks compared to the baseline, whereas the P300 was modulated by congruence only in mostly incongruent blocks and not in the baseline condition. This interaction between the component (N2/P300) and the condition (mostly incongruent/baseline) suggests that the strategic adaptation to conflict impacts both early conflict detection processes as reflected by the N2 component (van Veen and Carter, 2002; Nieuwenhuis et al., 2003; Folstein and Van Petten, 2008) and late strategic processes indexed by the P300 (Polich, 2007; Frühholz et al., 2011). This finding is noteworthy as it highlights that even early processes can be affected by top-down strategic control. It is in the agreement with the conflict-monitoring theory (Botvinick et al., 2001, 2004), according to which the anterior cingulate cortex (ACC) activation related to conflict detection and indexed in EEG by the N2 (van Veen and Carter, 2002) should be lower when incongruent trials are frequent (Carter et al., 2000; Kerns et al., 2004).

Unconscious processing of the primes was assessed by combining subjective reports and an objective forced-choice discrimination task. We included two types of blocks in this discrimination task, in order to mimic the trial structure used in the main experiment: one block contained only masked trials whereas the second contained 70% unmasked trials and 30% masked trials. Interestingly, the d' scores obtained in the block with both masked and unmasked trials were higher than those obtained in the block with only masked trials. This result could reflect either an increased conscious visibility or an increased unconscious objective performance and is reminiscent of previous studies on the influence of temporal and spatial attention and stimulus expectations on unconscious processing (Kentridge, 1999; Naccache et al., 2002; Ansorge and Neumann, 2005; Kiefer and Brendel, 2006; Marzouki et al., 2007; Kiefer, 2012) and on improved conscious visibility of stimuli presented at threshold (Sergent et al., 2013). Indeed, in the block with both masked and unmasked trials, subjects had strong expectations about the masked stimuli, which probably facilitated their perception. In the present study, we did not collect detailed subjective visibility assessment for each trial, so we

cannot disentangle between the hypothesis of an increased conscious visibility and the hypothesis of an increased unconscious objective performance. However, the larger Stroop effect at null d' in the masked trials included in unmasked blocks strengthens the second hypothesis, as one should expect less Stroop effect due to strategic processing for visible trials.

Interestingly, Jáskowski et al (2003) showed significant blockwise adaptation to conflict in masked trials, contrary to the results of the present study. They explained their results by a meta-cognitive process: they assumed that subjects were aware of the consequences of the high proportion of incongruent trials (higher number of errors) and regulated their behavior accordingly, even though they were not aware of the conflict itself. This interpretation fits with the idea that unconscious stimuli cannot trigger cognitive control directly, but that the conscious consequences of these stimuli can (Dehaene and Naccache, 2001; Kinoshita et al., 2008; Van Den Bussche et al., 2008; Desender et al., 2014). Note that the priming effects in Jáskowski et al's study are very strong (about 100 ms difference in reaction times and about 15% error rate in incongruent trials), allowing subjects to notice changes in these parameters. The results of our study could be explained in this meta-cognitive framework. In our experiment, the Stroop effect is small (e.g. for the masked baseline, mean = 5.5 ms) with low error rate for incongruent trials (e.g. for the masked baseline, mean = 5.6%), so it might be hypothesized that participants did not notice consciously the difference between congruent and incongruent masked trials and thus could not modulate their behavior in these trials. However, in our study, the instructions were explicit but no modulation according the proportion of incongruent masked trials was observed, suggesting that knowing consciously the structure of the task is not sufficient for subjects to adapt their behavior in masked trials and that this information should be extracted from responses themselves. Note that, in our study, the modulation of the Stroop effect requires semantic processing of the prime, whereas the paradigm used by Jáskowski et al requires only lower-level processes to identify the direction of an arrow, which could explain the observed differences in the amplitude of the priming effects (van Gaal et al., 2008; Dehaene and Changeux, 2011).

Moreover, the analysis of the ERPs revealed a modulation of the P300 component by congruence in masked trials included in mostly unmasked blocks, but not in masked trials included in fully masked blocks. Interestingly, in unmasked trials, this modulation was observed in mostly incongruent blocks, but not in the baseline condition, suggesting that it reflects a process involved in the strategic use of the prime. Therefore the P300 modulation observed in masked trials included among unmasked trials could be explained in terms of cognitive control triggered by conscious stimuli. The second ERP finding which is surprising in the current study is that the modulation of the N2 component by congruence observed in the unmasked baseline is not present in the masked baseline, although the behavior shows a differential processing of congruent and incongruent trials in this condition. This result might be explained by the relatively low number of trials in the baseline condition as well as by the low amplitude of behavioral effects in masked trials. It would thus be interesting to replicate the present results in an experimental paradigm triggering stronger behavioral effects, such as the one used by Jáskowski et al (2003).

Finally, we found that the use of the instructed strategy depended on the previous trial congruence and visibility, suggesting an online evaluation of this rule from trial to trial. First, following an unmasked congruent trial, the Stroop effect was similar to the one observed in the baseline condition and the modulation of the P300 component by congruence, which indexed the application of the instructed strategy, was not present. Second, following a masked trial, independently of its congruence, the Stroop effect in unmasked trials was similar to the one observed in the baseline condition, and no P300 modulation by congruence was observed. Note that the modulation of the P300 observed in masked trials intermixed with unmasked trials (see above) do not seem to influence the processing of subsequent unmasked trials. This may be explained by the evanescence of masked effects (Greenwald et al., 1996; Dupoux et al., 2008). Finally, we showed that all these results could not be explained by physical repetitions across trials (Mayr et al., 2003). These trial-to-trial differences suggest that the instructed strategy is evaluated on each trial based on the recent history, as the Stroop effect was strongly reduced

relative to baseline only after incongruent trials, which are valid with the strategy. Therefore, these differences may be described as the dynamics of 'subjective strategy adherence', which were observed exclusively for unmasked trials. This result may suggest that changes in subjective strategy adherence require conscious access to the stimuli. An important issue would be to measure reportability of such strategy adherence, as this process could require conscious access to stimuli without being reportable itself. It would thus be interesting to test how much subjects adhere to the instructed strategy, by assessing their confidence in the strategy on each trial in future experiments. The modulation of strategy transfer to masked trials by the strength of conscious cognitive control processes as indexed by the trial-to-trial adherence to the strategy should also be tested.

7. Acknowledgments

This work was supported by the program 'Investissements d'avenir' ANR-10-IAIHU-06, an AXA Research Fund grant to IEK and an 'Equipe FRM 2010' grant of Fondation pour la Recherche Médicale (FRM) to LN.

CHAPTER 4. COGNITIVE DISSONANCE RESOLUTION IS RELATED TO EPISODIC MEMORY

1. Presentation of the article

This third study was designed to test how subjects deal with inconsistencies in their own behavior. We addressed this question within the context of the cognitive dissonance theory and used the free-choice paradigm. Subjects were thus asked to rate vacation destinations, then they had to choose between destinations they rated equally and finally they had to rate the destinations again. Only half of the destinations were used in this condition. The other half was used in a control condition allowing us to control for the critical statistical artefact highlighted by Chen and Risen (2010), the 'rate-rate-choose' condition. Moreover, we included a memory test at the end of the experiment to test the role of explicit memory of the choice in choice-induced preference change. This study allowed us to test the way subjects adapt their behavior when facing inconsistencies in their own behavior and whether these inconsistencies need to be accessed consciously, through memory.

2. Abstract

The notion that our past choices affect our future behavior is certainly one of the most influential concepts of social psychology since its first experimental report in the 50s, and its initial theorization by Festinger within the "cognitive dissonance" framework. Using the free choice paradigm (FCP), it was shown that choosing between two similarly rated items made subjects reevaluate the chosen items as more attractive and the rejected items as less attractive. However, in 2010 a major work by Chen and Risen revealed a severe statistical flaw casting doubt on most previous studies. Izuma and colleagues (2010) supplemented the traditional FCP with original control conditions and concluded that the effect observed could not be solely attributed to this methodological flaw. In the present work we aimed at establishing the existence of genuine choice-induced preference change and characterizing this effect. To do so, we replicated Izuma et al.' study and added a new important control condition which was absent from the original study. Moreover, we added a memory test in order to measure the possible relation between episodic memory of choices and observed behavioral effects. In two experiments we provide experimental evidence supporting genuine choice-induced preference change obtained with FCP. We also contribute to the understanding of the phenomenon by showing that choice-induced preference change effects are strongly correlated with episodic memory.

Key Words: Cognitive dissonance, Choice-induced preference change, Conflict, Attitude, Values, Memory

3. Introduction

In 1957, Festinger coined the expression 'cognitive dissonance' to convey a general theory postulating that individuals strive to decrease the discomfort generated by conflicting cognitions by modifying one's cognitions (Festinger, 1957). Since then, this concept has intensively stimulated empirical and theoretical research in social psychology, and it has been so widely popularized that it is frequently used in mass-medias. From an experimental perspective, cognitive dissonance is usually tested through the five following paradigms: induced-compliance (Festinger and Carlsmith, 1959), belief-disconfirmation (Festinger et al., 1956), effort justification (Aronson and Mills, 1959), misattribution (Fries and Frey, 1980) and free-choice paradigms (Brehm, 1956). The free-choice paradigm (FCP) originated a central claim of cognitive dissonance theory, namely, that our past choices affect our future behaviors, preferences and beliefs.

3.1. An intense theoretical debate about the Free-Choice Paradigm (FCP)

A typical FCP experiment is composed of three blocks. First, subjects rank or rate items according to their desirability (e.g.: food items, car models, holiday destinations, etc…). Ranking and ratings are used interchangeably in FCP paradigms; from this point on we will use only the latter. Second, they are engaged in a forced-choice task during which they have to choose between two closely rated items. Finally, they perform a second rating on the same items. Choice-induced preference change is defined by a tendency to increase ratings of chosen items, and to decrease those of rejected items. This 'spreading of alternatives', or 'spread', is the hallmark of the phenomenon, and has been considered as diagnostic of preference change. This paradigm has been extensively used during the last decades. However, there is still an intense theoretical debate about the precise mechanisms subtending choice-induced preference change. Schematically, the literature contains two main and opposite types of theoretical accounts.

On the one hand, many different models call for high-level cognitive, self-related (see(Egan et al., 2010)) or metacognitive top-down accounts of choice-induced preference change. This class of models includes the original Festinger's model of cognitive dissonance theory (Festinger,

1957), as well as other reformulations of it (for extensive review see (Cooper, 2007)). Cognitive dissonance postulates the existence of a succession of executive stages triggered by the choice: cognitive conflict generation, detection, monitoring, and resolution. Similarly, other models departing from cognitive dissonance theory share this high-level view of choice-induced preference change. Self-consistency theory proposes that the spread is driven by the need for subjective consistency between previous actions (choice) and current explicit ratings (Aronson, 1968; Thibodeau and Aronson, 1992). Self-perception model interprets the spread as reflecting the dynamics of an internal model of the self who is trying to account for his own behavior in the absence of any direct access to an encapsulated value system (Bem, 1972). Recent neuroimaging findings implicating regions of the executive network such as the anterior cingulate cortex (ACC) strengthen this class of models (van Veen et al., 2009).

On the other hand, other empirical studies suggest that choice-induced preference change is a low-level process which may occur unconsciously and independently from episodic memory and executive control. This conception of choice-induced preference change is supported by the report of spread in amnesic patients (Lieberman et al., 2001), in normal controls with very long delays between the two ratings (Sharot et al., 2012), in young infants and even in Capucin monkeys (Egan et al., 2007, 2010). In the same line, findings of choice-induced preference change correlates in the activity of sub-cortical regions included in the motivation network, such as striatal regions (Sharot et al., 2009; Izuma et al., 2010) seem congruent with this conception. Indeed, finding a correlate of the spread in the neural value system supports that it is a direct marker of value modifications.

3.2. A statistical artefact potentially flawing most FCP studies

Within this scientific debate, Chen and Risen reported in 2010 that spread could be observed in the FCP even under conditions of stable preferences (Chen and Risen, 2010). Their claim relied on two assumptions. First, they presumed that ratings are noisy measures of subjects' preferences. Second they suggested that subjects' choices give additional information about their preferences. Accordingly, if two items, A and B, were similarly rated by a subject, and he

chooses A over B, then the most probable account would be that item A is actually preferred over item B, and that initial identical ratings corresponded respectively to under-estimation and over-estimation of genuine preferences for A and B. Therefore, as a result of regression to the mean, one could expect an increase of rating for item A and a decrease of rating for item B during the second rating, without any change in preferences for these items. In other terms, spread is inherent to the FCP design. This devastating theoretical claim was experimentally confirmed by a clever design adding to the classical 'rating-choice-rating' (RCR) design, a new 'rating-rating-choice' (RRC) sequence. Chen and Risen showed that a spread could be obtained even with this latter sequence, in which choices were made after the second rating. Crucially, in their study RRC and RCR spreads did not differ, suggesting that genuine choice-induced preference change is either inexistent or in any case much smaller than initially reported since 1956. Izuma and Murayama ran simulations of the FCP under Chen and Risen' assumptions, and confirmed this statistical flaw (Izuma and Murayama, 2013). They recommended that any attempt to demonstrate a choice-induced preference change should control for it. Since Chen and Risen' major study, the inclusion of such a RRC sequence in experiments using the FCP seems to be a prerequisite for interpreting any spread as an evidence of genuine choice-induced preference change. It is central to note that most experimental studies using the FCP performed before 2010 did not control for this major artefact. The first studies controlling for it revealed much smaller effect sizes than in the pre-2010 original studies(Izuma and Murayama, 2013). Therefore, the mentioned theoretical debate remains open given the fragility of previous empirical findings.

In response to Chen and Risen' criticism, several attempts were made to validate the spreading of alternatives obtained with FCP. Alós-Ferrer & Shi formally refuted the mathematical model presented by Chen and Risen but agreed that on some occasions a positive spread could be observed in the absence of any preference change, and confirmed that 'the fact that expected spreading for specific rating distances and model specifications might be non-zero makes improved experimental design very valuable' (Alós-Ferrer and Shi, 2012). Moreover, a few

studies have attempted to control for this artifact and still observed choice-induced preference change effects. In particular, Johansson and colleagues (2013) used the choice blindness paradigm to manipulate choices and found a spread effect (Johansson et al., 2013). Sharot et al. (2012) used a RCR/RRC design to probe very long-lasting effects (>2.5 years). They found a significantly higher spread for RCR than for RRC, but by definition the choice-rating delays were highly asymmetrical between the RCR condition (both rating-choice delays were short) and the RRC condition (the delay between the first rating and the choice was very long >2.5 years). This asymmetry may raise an issue to interpret spread values. Three other recent studies used an ingenious blind choice condition preventing subjects' choices from revealing more information about their real preferences (Egan et al., 2010; Sharot et al., 2010; Nakamura and Kawabata, 2013). However, these results are not immune from criticism. Indeed, the methodology used by Egan et al. (2010) to test infants and monkeys has been put into question. Indeed, in infants, attitude change was estimated by a second blind choice, which thus does not reflect preferences. Moreover, in monkeys, attitude change was assessed by 10 open choices, following the critical blind choice. But according to cognitive dissonance theory, these choices should influenced each other in turn, so the results are hard to interpret (see (Risen and Chen, 2010; Holden, 2013; Izuma and Murayama, 2013) for detailed reviews of this study). Moreover, Sharot et al.'s study (2010) (Sharot et al., 2010) reported a significant spread only for chosen items while the canonical spread was marginal (p=0.1).

3.3. Focus on the Izuma et al. (2010) study using a RCR and RRC design

Finally an interesting study by Izuma et al. (2010) used the design recommended by Chen & Risen and reported a significantly larger spread in the RCR condition than in the RRC condition. Interestingly, this work is both cited by proponents of the 'high-level' (Kitayama et al., 2013; Mengarelli et al., 2013) (however none of these two studies properly controlled for the Chen & Risen's artifact) and of the 'low-level' views (e.g.: (Sharot et al., 2012)). A key reason for this ambiguity is to be found in the detail of the design used by Izuma and colleagues.

Indeed, during the second rating of the RCR sequence, each item was labeled with a reminder of the choice the subject had previously made on that item (e.g.: '*you rejected it*'). Obviously, such a reminder was delivered only in the RCR sequence, given that in the RRC sequence the choice was made after the second rating. As a consequence, Izuma et al. did not contrast RCR with RRC conditions, but they rather compared 'RCR + explicit choice reminder' versus RRC condition. The probable reason for not choosing the minimal contrast (RCR vs RRC) was to maximize chances of observing choice-induced preference change by capitalizing on all possible cognitive mechanisms (high and low levels). As such, Izuma et al. study succeeded in providing a significant behavioral effect. However, the asymmetry between RCR and RRC conditions makes it simply impossible to disentangle between these distinct cognitive mechanisms.

In the present work we replicated Izuma et al. study and added a strict RCR (without reminder cue) condition. Moreover, we added a memory test in order to measure the possible relation between episodic memory of choices and observed behavioral effects. In a set of two different experiments we reveal a novel and strong correlation between episodic memory of choices and choice-induced preference change.

4. Experiment 1

4.1. Methods

4.1.1. Ethics Statement

These experiments have been approved by the Pitié-Salpêtrière ethical committee. All the 65 subjects gave their written informed consents, and were paid 10 Euros to participate in the experiment. All investigations were conducted according to the principles expressed in the Declaration of Helsinki.

4.1.2. Participants

Twenty-six subjects were included in the 'Reminder' group (17 women; age M = 25.3 years old, SD = 5.6; 92% right-handed), and 25 in the 'No reminder' group (16 women, age M = 26.4, SD = 5.2; 96% right-handed). They reported normal, or corrected-to-normal, visual acuity. Data from three participants were excluded (2 from the 'Reminder' group, and 1 from the 'No reminder'). One dataset was not saved due to technical failure. The remaining 2 excluded participants essentially used the highest rating values within the 8-values scale in rating 1 (rating 1 median = 8), with a ceiling effect biasing the possible rating changes only to decreased values.

4.1.3. Stimuli

Stimuli were 120 colored images of potential vacation destinations. Image subtended 5.3º of visual field. Destination name was printed (font size=30) below the image. In 'Rating' blocks, one image appeared in the center of the screen in each trial, whereas in 'Choice' blocks, two targets were presented 4.8º off-center, to the left and to the right, in each trial.

4.1.4. Procedure

In the 'Reminder' group, we followed a procedure similar to the one reported by Izuma et al. (2010) (Izuma et al., 2010). The experiment was composed of five blocks: two 'Rating' blocks (Rating 1 and 2) and two 'Choice' blocks (Choice 1 and PostEx Choice) and a 'Memory' block for the 'No Reminder' group (Fig. 4.1).

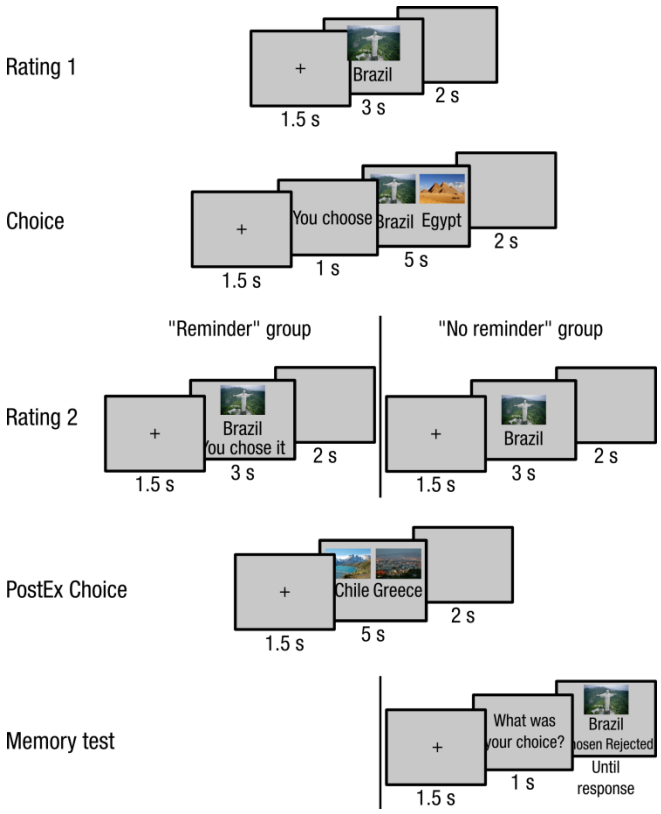

FIGURE 4.1: EXPERIMENTAL PARADIGM FOR THE 'REMINDER' AND THE 'NO REMINDER' GROUPS

The experimental paradigm included 5 stages: Rating 1, Choice 1, Rating 2, Choice 2 and Memory test. The procedure differed between the 'Reminder' and 'No reminder' groups in Rating 2 and Memory test. In Rating 2, previous choice (either subject's or computer) was indicated for the 'Reminder' group (left panel) but not for the 'No reminder' group (right panel). Additionally, subjects in the 'No reminder' group performed a memory test designed to check whether they have remember the information corresponding to the reminder present in Rating 2 for the 'Reminder' group.

The experiment started with 'Rating 1' block, which included 120 trials. Each trial began with a fixation point presented during 1.5-seconds. Then one vacation destination was centrally presented for 3 seconds, and followed by a blank screen lasting for 2 seconds during which subjects were requested to report how much they would like to spend their vacation in this destination using an eight-point scale (1 = 'I do not want to go there at all', 8 = 'I definitely want to go there'). Subjects responded using the 1-8 number pad buttons of a regular keyboard. In the absence of response, the experiment proceeded to the next trial.

This first block was followed by 'Choice 1' block. In this block, every trial started with a fixation point presented during 1.5-second. Then a task instruction appeared for 1 second, indicating whether the subject should choose between the two destinations to be proposed on the next screen ('You choose'), or whether he should observe computer's choice ('Computer chooses'). Then, two destinations were presented side-by-side for 5 seconds. Only in the 'computer chooses' trials a black frame surrounded the chosen destination. Trials ended with a 2-second presentation of a blank screen. Subjects had to report manually either computer's choice or their own choice using the left/right arrows keyboard keys. Crucially, destinations were coupled according to subjects' responses in 'Rating 1' block, as to create 3 experimental conditions: 'Self-Easy', 'Self-Difficult' and 'Computer'. 'Self-Easy' trials were composed of a highly rated destination (rating ≥ 5) and a low rated destination (rating < 5) that differed in rating by at least 3 points. 'Self-Difficult' and 'Computer' trials were composed of two highly rated destinations (rating ≥ 5) that differed in rating by no more than one point. The number of couples for each sequence was equal, but changed from subject to subject (from 8-19 for each sequence).

The third experimental stage was 'Rating 2' block, and differed between the 'Reminder' and 'No reminder' groups. In the 'No reminder' group, this block was strictly identical to 'Rating 1'. In the 'Reminder' group, a written reminder indicating subject's or computer's choice concerning this item during 'Choice 1' block (e.g.: 'you chose it' or 'computer rejected it') was added. As in Izuma et al. (2010), subjects were instructed to ignore these comments.

This 'Rating 2' block was followed by 'PostEx Choice' block, in which only couples of destinations used in the 'Computer' condition during 'Choice 1' block were presented. Subjects had to choose between these items, as described for 'Choice 1' block.

Finally, a memory test was added at the end of the experiment for the 'No reminder' group. Every trial began by an instruction presented for 1 second and specifying which event should be recalled on the following destination: either subject's own choice or computer's choice concerning this item in 'Choice 1' block. Note that this memory test was addressing the information reminded to subjects during 'Rating 2' in the 'Reminder' group. Importantly,

subjects were not informed about the existence of this final memory test when they performed the first 4 blocks, so they were not explicitly instructed to memorize their choices.

As we were interested in testing whether subjects remembered their choices, we considered items as remembered only if subjects correctly reported whether they had chosen or rejected each of the two coupled items.

4.2. Results

We first replicated results reported by Izuma et al (2010 - see Fig. 4.2a and 4.2b). Indeed, within the 'Reminder' group which is a direct adaptation of Izuma et al.'s procedure to our vacation destinations dataset, preference for destinations which were rejected in the Self-Difficult condition significantly decreased compared with rejected destinations in the Self-Easy condition ($t(23) = 9.67$, $p<10^{-8}$) or those rejected in the Computer condition ($t(23) = 4.14$, $p<0.001$). These comparisons were used before Chen and Risen's criticism to assess the existence of preference change, but they do not allow controlling for the statistical artefact highlighted by Chen and Risen (Chen and Risen, 2010). Importantly, in this group, we also observed a significant difference between the critical condition and the proper control condition: the spread was significantly larger for the Self-difficult/RCR sequence than for the PostEx-Choice/RRC sequence ($t(23)= 5.1$, $p<0.001$). This replicates the critical behavioral result reported by Izuma et al (Izuma et al., 2010).

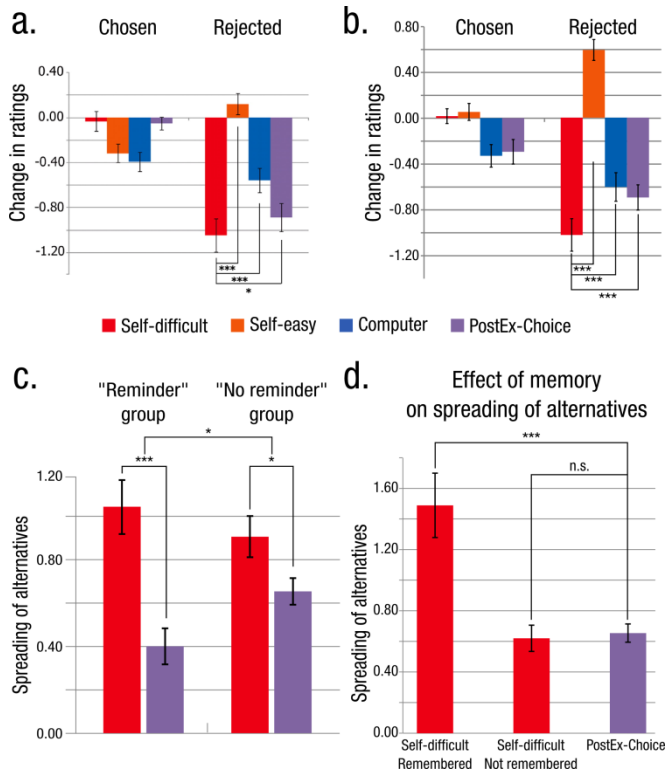

FIGURE 4.2: SPREADING OF ALTERNATIVES IN EXPERIMENT 1

(a) *Results obtained by Izuma et al. 2010 (adapted from PNAS). Bars represent the change in ratings between Rating 2 and Rating 1 for chosen items (left) and rejected items (right) in each condition. The critical comparison (corresponding to the comparison between RCR and RRC sequences respectively) is between Self-Difficult and PostEx-Choice. * p< 0.05, *** p<0.001 (paired t-test, one tailed).*

(b) *Replication of Izuma et al's experiment ('Reminder' group). *** p < 0.001 (paired t-test, one tailed).*

(c) *Spreading of alternatives in the 'Reminder' (left) and 'No reminder' (right) groups. The interaction between the sequence (RCR – Self-difficult/RRC – PostEx-Choice) is significant (* p < 0.05, ANOVA) and the critical comparison between RCR - Self-difficult and RRC - PostEx-Choice conditions is significant in both groups (* p < 0.05, *** p < 0.001, paired t test, two tailed).*

(d) *Effect of memory on spreading of alternatives. The difference between the spread in the RCR – Self-difficult and the RRC – PostEx-Choice was significant only in pairs for which the choice was remembered (*** p < 0.001, paired t test, two tailed).*

We then focus on the two critical conditions: the Self-Difficult/RCR condition and the Post-Ex Choice/RRC condition, as the Self-Easy and the Computer are control conditions which do not take into account Chen and Risen's criticism. We ran an analysis of variance (ANOVA) on the spread observed after difficult choices with an inter-subject factor ('Reminder'/'No reminder')

and a within-subject sequence factor (Self-Difficult RCR/Post-Ex Choice RRC). The reminder factor was not found significant ($F<1$). A main effect of sequence was observed ($F(1,92)=28.5$; $p<0.001$) with larger spread for RCR than for RRC sequence. Crucially, these two factors interacted ($F(1,92)=5.5$; $p=0.02$) with larger difference in spread between RCR and RRC sequences in the 'Reminder' group than in the 'No reminder' group (mean difference in spread between RCR and RRC sequence for 'Reminder' group: 0.64, SEM = 0.13, for 'No reminder' group: 0.25, SEM = 0.11). Restricted analyses showed that in each of the two groups, spread was significantly larger for RCR than for RRC (for the 'Reminder' group: $t(23)= 5.1$; $p<0.001$ and for the 'No reminder' group ($t(23)= 2.3$, $p=0.03$). Taken together these results suggest the existence of genuine choice-induced preference change, even after taking into account both the statistical flaw revealed by Chen and Risen (Chen and Risen, 2010) and the confound introduced by the reminder. They also reveal the strong impact of the reminder, suggesting that the availability of choice information during the second rating enhances preference change.

In order to refine this new finding, we focused on the 'No reminder' group in which choice-information was not delivered during the second rating. Overall, while subjects were at chance for remembering computer' choices, they were still above chance for their own choices ($t(23) = 9.87$, $p <0.001$), leaving open the possibility that the availability of choice information during the second rating was modulating the spread. We then categorized each item as correctly or incorrectly memorized using the memory test performance. Crucially, remembered RCR trials showed a larger spread than RRC trials ($t(23) = 3.72$, $p = 0.001$), whereas no such a difference could be observed on forgotten RCR trials ($t <1$).

Taken together these results reveal the importance of the availability of choice information in the magnitude of preference change effects, be it through an external reminder or a spontaneous recall of the choice.

5. Experiment 2

We designed and ran a second experiment with the following four objectives.

First, we aimed at replicating in a new group of subjects the effect of choice memory on the spread.

Second we modified the memory task in order to be able to cross RCR/RRC sequences with remembered/forgotten status of individual items, and therefore to test for the existence of an interaction between these two factors. Indeed, in the 'No reminder' group of the first experiment, the memory test was designed to probe the availability of information related to the reminder used by Izuma et al. (2010) (Izuma et al., 2010). Therefore, the memory test probed subjects' and computer's choices from the first choice block, but not subjects' choices from the second choice block. So, this design could not allow us to properly cross memory and RCR/RRC factors, given that we did not have memory information on the choices made in the RRC sequence.

Third, one issue raised by our findings relates to the timing of preference change. Does it happen during the second rating, or could it occur immediately after the choice? As is, our memory test does not allow us to solve this issue given that a late recall of the choice is not diagnostic of the dynamics of preference change. In order to address this question, we compared spread values between blocks similar to the 'No reminder' group of experiment 1, and blocks during which subjects had to perform a demanding 2-back task immediately after their choices. This manipulation aimed at interfering with working memory and conscious elaboration about subjects' own preferences during the choice.

Finally, we also aimed at disentangling memory from subjective relevance. Indeed, memory performance observed in experiment 1 could potentially be a side-effect of subjective relevance: subjects remembering their favorite or highly-rated choices better. Under this hypothesis, memory would not play a direct role in preference change. To address this issue, we coupled equally rated items from the whole rating scale and not only from the high rating values as we did in experiment 1.

5.1. Methods

5.1.1. Participants

Thirty nine subjects (22 women; age M = 22.9 years old, SD = 3.1; 95% right-handed) participated in this study (see Ethics Statement above). All reported normal or corrected-to-normal visual acuity. The data from 3 participants were excluded from the analysis. These subjects were removed because their responses were shifted toward scale's extremities ('Rating 1' median=8 for two of them and 1 for the other).

5.1.2. Stimuli

Stimuli were the same images of vacation destinations as in the experiment 1. In the n-back task, stimuli were capital letters (size 18) presented at fixation for 1 second.

5.1.3. Procedure

Procedure was similar to the one used in the first experiment with four major differences (see Fig. 4.3a). First, subjects' responses were not limited in time. Second, items were coupled by sorting them according to their first rating resulting in 60 pairs of similarly rated items. Third, each 'Choice' block was divided into 2 sub-blocks composed of 15 couples of items. Pairs of items were randomly attributed to one of these sub-blocks. The first sub-block was identical to the 'Choice 1' block in the previous experiment. In the second sub-block, subjects had to perform a 2-back task after each choice. Ten letters were presented at fixation, and subjects had to report a letter repetition if it was separated by two other letters by pressing the spacebar key. Fourth, in the memory test, subjects had to report whether they had chosen or rejected the item during the first or the second 'Choice' block.

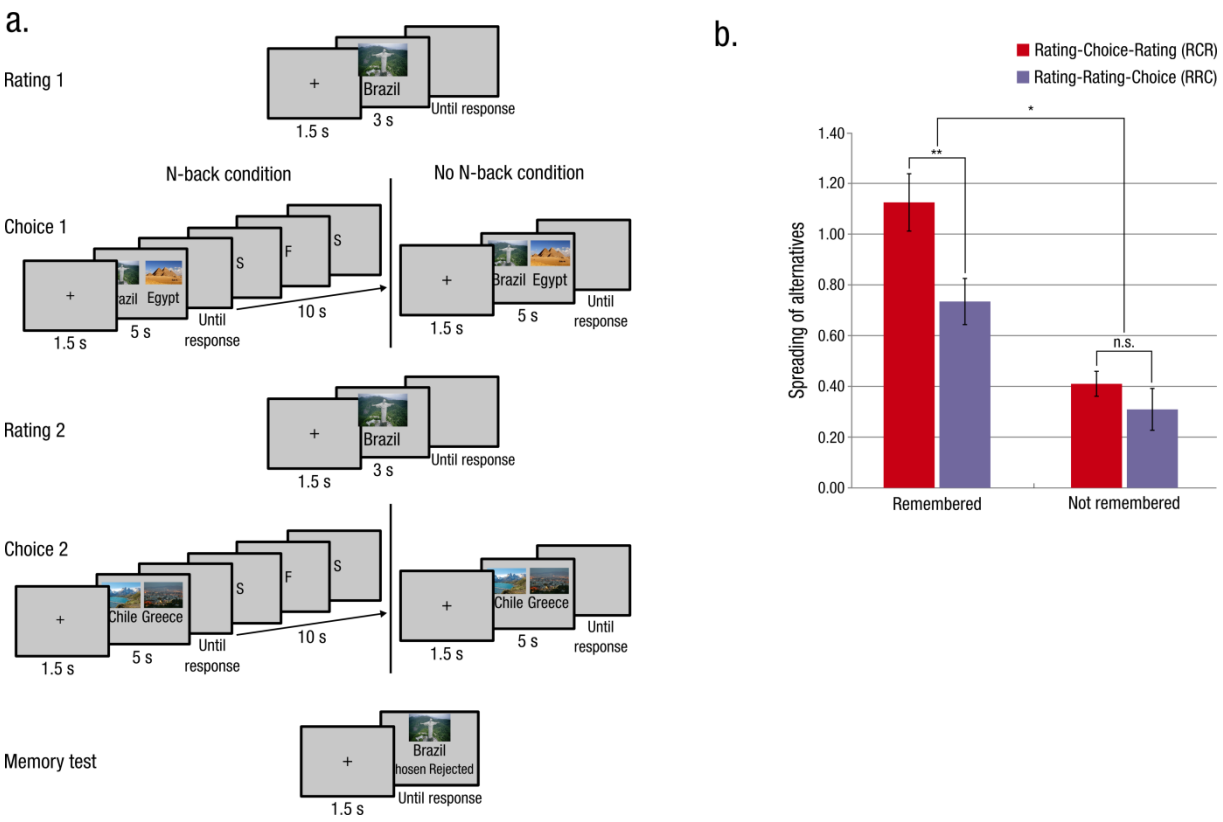

FIGURE 4.3: EXPERIMENTAL PARADIGM AND EFFECT OF MEMORY IN EXPERIMENT 2

(a) *Experimental paradigm for Experiment 2. Five stages were included: Rating 1, Choice 1, Rating 2, Choice 2 and Memory test. Each 'Choice' block comprised two sub-blocks: one in which the choice was immediately followed by a 2-back task and one including only the choice.*

(b) *Effect of memory on spreading of alternatives. The interaction between the sequence (RCR/RRC) and the memorization of the choices was significant (* p = 0.019). The difference between the spread in the RCR and the RRC sequences was significant only in pairs for which the choice was remembered (** p < 0.01, paired t test, two tailed).*

5.2. Results

We first tested the effect of the manipulation we used by an ANOVA on spread including sequence (RCR/RRC) and manipulation (N-back/No N-back) as factors. A main effect of sequence was found ($F(1,140) = 5.54$, $p = 0.02$). The main effect of manipulation almost reached significance ($F(1,140) = 3.68$, $p = 0.06$). But the interaction between sequence and manipulation was not significant ($F<1$).

Memory-test scores showed that the N-back manipulation was effective as choices that were followed by it were remembered less than those that were not (mean N-back: 0.32, SEM = 0.02, No N-back: 0.44, SEM = 0.03; $t(35)=4.95$ $p<0.001$).

As we were mainly interested in the relationship between memory and spread we also tested the effect of memory in this task using a linear mixed-effects model including the spread for each pair of items for each subject, with the 'Sequence' (RCR/RRC) and the 'Memory' (remembered/forgotten) factors defined as fixed factors and subjects defined as a random factor. Indeed, as opposed to a classical ANOVA, such a model does not require prior averaging of the spread for each subject and thus offer the possibility to handle the heteroskedasticity related to the unbalanced number of items in each condition (Baayen et al., 2008). Significance of the fixed effects was assessed using the Kenward-Roger approximation for degrees of freedom of the denominator, with the 'lmerTest' package in R. A main effect of memory was found ($F(1,2112) = 64.7$, $p < 10^{-10}$), as well as a main effect of sequence ($F(1,2124.3) = 15.7$, $p < 10^{-4}$). The critical interaction between memory and sequence was significant ($F(1,2145) = 5.5$, $p = 0.019$ – Fig. 4.3b). Indeed the difference between RCR and RRC was highly significant for remembered items ($F(1,795.6) = 15.3$, $p<10^{-4}$ – Fig. 4.3b), but was not significant for forgotten items ($F(1,1325.4)=1.68$, $p=0.195$ – Fig. 4.3b). Whereas we observed, as we expected, a larger spread for remembered than for forgotten items in the RCR sequence ($F(1,1021.9) = 53.4$, $p < 10^{-10}$), the same effect was surprisingly also found in the RRC sequence ($F(1, 1076) = 17.0$, $p < 10^{-4}$).

6. Discussion

In this study, we obtained four results that will structure our discussion.

First, we confirmed the existence of the artifact discovered by Chen and Risen (Chen and Risen, 2010). Indeed, spreads obtained in the RRC control sequence were systematically significantly larger than zero. This confirms Chen and Risen' argument that choices hold information about preferences and that they should not be only considered as a source of preference change.

Moreover, we found a significantly larger spread for items which were correctly remembered as chosen or rejected in the RRC sequence than for items which were not correctly remembered. Within the Chen & Risen framework, this unexpected effect could simply reflect the impact of a better attentional engagement during some choices, which would in turn reveal more accurate information about preferences. Indeed, given that memory encoding depends on attention (see (Chun and Turk-Browne, 2007) for a review), a choice that was later remembered was presumably processed with more attention than a choice that was later forgotten. Moreover, attended choices may well provide more information about subjects' preferences than less attended choices. As a direct consequence of this bias and in agreement with our results, larger spread would be observed for memorized RRC items than for forgotten ones. This would also be reflected in an increased number of pairs in which spread was positive. Our results did show this pattern, as the mean proportion of couples for which positive spread was achieved in the RRC sequence was higher ($t(35)=2.51$, $p<0.02$) for remembered choices (mean=0.842, SEM=0.020) than for forgotten choices (mean= 0.774, SEM=0.025).

Second, in spite of the artifact pointed out by Chen and Risen, we replicated the finding reported by Izuma et al. (2010), by observing a significantly larger spread in the RCR sequence than in the RRC sequence. This result, which we replicated in our two experiments, strengthens the existence of genuine spreading of alternatives in the FCP paradigm. To date, our current work counts among the rare robust results within an empirical literature weakened by the seminal study of Chen & Risen (see also Johansson et al., 2013; Nakamura and Kawabata, 2013).

Third, we showed that the reminder cue used by Izuma and colleagues (Izuma et al., 2010) affected drastically the measured spread. As we mentioned above, Izuma et al. did not compare RCR versus RRC, but rather 'RCR + explicit choice reminder' versus RRC. We observed that in the absence of such a reminder cue, the difference between RCR and RRC spreads was much smaller than when a cue was used. This result is important in regard to the current theoretical debate about the precise mechanism at work in choice-induced preference change. According to models postulating automatic updating of values independent of episodic memory, such an explicit cue

should not affect spread values. Alternatively, self-related models postulate that such episodic memory cue should decisively affect the spread by loading conscious working memory with choice information. Our results thus tend to favor this second hypothesis.

Finally and most importantly, we found a strong relation between choice memory and choice-induced preference change. Only remembered items showed a significantly larger spread for RCR than for RRC. We replicated this original finding across two experiments with different subjects and two different paradigms. It may explain, in part, the boosting effect of the reminder cue used by Izuma et al (Izuma et al., 2010) on preference change. In experiment 1, we showed that in the absence of the reminder cue, a spread was observed in pairs of items, for which choice was remembered, and was absent in pairs for which choice was forgotten. In experiment 2, we tried to manipulate choice encoding by loading working memory and attentional resources, using a 2-back task immediately after choice in half of the trials. We succeeded only in part, as choices that were followed by the 2-back task were mostly forgotten, but choices that were not followed by the task were not all remembered. We then examined spreading of alternatives according to subjects' performance on the memory test, irrespective of the post-choice manipulation. This analysis confirmed that the spread was larger for RCR sequence than for RRC sequence only if choices were remembered.

This correlation between spread and memory raises the issue of causality between episodic memory of choices and spread, and of the timing of observed preference changes. Is this correlation pointing to a causal link between episodic memory of choices and spread? One possible interpretation of our results, - which supports self-related theories -, posits that preference change occurs only when subjects are confronted to a self-coherence context such as the second rating (R2) for items which were correctly remembered as chosen or rejected. Under this assumption, preference change would not occur implicitly or unconsciously immediately after the choice. In other words, choice-induced preference change may actually be a 'choice-rating consistency induced' preference change occurring exclusively during the second rating. Alternatively, supporters of implicit theories of preference change may propose that memory of

choices could be more accurate for items showing a large spread than for those with more stable ratings. Under this view, a reverse causality would be proposed: memory would not cause preference change, but preference change would enhance performance in the memory test. This would be coherent with the observation of significant spread in amnesic patients before Chen and Risen's criticism ((Lieberman et al., 2001)). Such an enhancement of performance in the memory test could stem either from a direct increase of episodic memory or from a guessing bias: subjects would guess their previous choices by probing their current updated value of a given item to infer their choice. Both the episodic memory enhancement and the guessing strategy hypotheses share a common prediction: this effect should be item-specific and not related to pairs of items proposed during the choice. Thus, we should observe a larger mean absolute value of ratings variation (mean of |R2-R1|) for remembered than for forgotten items. We ran this analysis and observed no difference of this index between remembered vs forgotten at the single item level in the RCR ($t(35)=-0.47$; $p=0.6$) and in the RRC ($t(35)=-0.34$; $p=0.34$) sequences. Moreover, our criterion of correct memory performance (correct answers to both items presented during a choice episode) was more prone to target episodic memory of the whole choice than episodic memory of a single item or choice guessing because both items of a given choice had to be correctly answered to categorize the memory answer as correct.

Taking all these considerations together, our findings seem to better support self-related models of preference change, even if the correlation we observed between memory performance and spread is far from being a direct evidence of a causal link.

However, we should improve the memory test in a way to disentangle the respective roles of episodic memory and inferential guessing. For instance, a subjective report such as a "remember/know" or "remember/guess" task could be added in order to better disentangle between these two possibilities. Additionally, the use of physiological measures or functional brain imaging recordings in the same paradigm could be useful to determine if preference change occurs before or after the second rating.

Another related question stemming from our results concerns the nature of the preference change effect: is it only a transient representation elicited by a need of consistency between past and present, within a working-memory system, or is it rather durably encoded in subject's value system? Future experiments could resolve these issues by examining the dynamics of preference change and the genuine impact of episodic memory on it. For instance, exploration of amnesic versus dysexecutive patients may help dissecting the role of episodic memory and executive functions in the processing of coherence at each experimental stage (choice and second rating). Functional brain-imaging could also be used to probe the respective contributions of episodic memory, executive control (van Veen et al., 2009) and value networks (Sharot et al., 2009) during both the choice and the second rating.

We conclude by stating that genuine choice-induced preference change effect does exist in the FCP, after controlling for two important confounds, and that it is related to episodic memory. Our results tend to support self-related models of preference change. But the precise role of conscious, - accessible to subjective reports -, and of non-conscious processes remain to be addressed in future studies.

7. Acknowledgments

We thank Mariam Chammat for her useful commentaries on this manuscript.

CHAPTER 5. WHEN DOES EPISODIC MEMORY AFFECT CHOICE-INDUCED PREFERENCE CHANGE?

1. Presentation of the study

This fourth study was also designed to assess the way subjects deal with inconsistencies in their own behavior. In the behavioral study presented in the previous chapter, we identified a crucial role of memory: choice-induced preference change was only observed when subjects remembered their choices. However, two questions could not be addressed by this behavioral study. First, was memory involved during the second rating? Second, when does choice-induced preference change occur? To address these issues, we designed a new fMRI experiment. We explored the neural underpinning of choice-induced preference change and specifically tested the activation of memory related brain areas during the second rating. Moreover, we included an incidental task between the first choice and the second rating in order to assess whether preferences had already change at this stage. Indeed, the brain valuation system is thought to be activated even when subjects are not performing a rating task, so we could evaluate which values were attributed to the different items at this stage. The analysis of these data is still ongoing, which is why only preliminary results are presented in this chapter.

2. Introduction

The idea that actions can influence, not just reflect, preferences has been highlighted by the theory of cognitive dissonance, proposed by Festinger (1957). According to this theory, when subjects hold two conflicting cognitions, they tend to reduce the dissonance caused by this conflict. A classic experimental paradigm supporting this theory is the free-choice paradigm (Brehm, 1956). In this task, subjects are first asked to rate items, such as vacation destinations or food items for example. Then they make choices between pairs of items. Finally, they rate all the items again. The typical results obtained in such an experiment show that for difficult choices, chosen items tend to be rated higher after the choice, while rejected items tend to be rated lower. This effect is termed spreading of alternatives. These results have been replicated several times since 1956 (e.g. Steele et al., 1993; Heine and Dehman, 1997). Moreover, the brain mechanisms related to dissonance reduction have been explored using functional magnetic resonance imaging (fMRI). The striatum and the ventro-medial prefrontal cortex (vmPFC) have been associated with choice-induced preference change (Sharot et al., 2009; Izuma et al., 2010; Jarcho et al., 2011; Qin et al., 2011; Kitayama et al., 2013), while the anterior cingulate cortex (ACC) and the insula have been implicated in the detection of the conflict induced by the choices (Izuma et al., 2010; Jarcho et al., 2011; Qin et al., 2011; Kitayama et al., 2013). Interestingly, similar brain areas have been implicated in cognitive dissonance resolution, using another experimental paradigm, the induced compliance paradigm (van Veen et al., 2009).

However, the behavioral results obtained with the free-choice paradigm have been challenged by the work of Chen and Risen (2010). According to these authors, in the free-choice paradigm, spreading of alternatives could be obtained without any change in real preferences. Their argument relies on two assumptions: first, subjects' ratings are noisy and thus are not perfect measures of one's true preferences; second, choices are at least partially guided by true preferences, and thus they reveal information about these true preferences. In difficult choices, subjects are presented with pairs of items they rated equally. But according to the first

hypothesis, this does not necessarily mean that the true preferences for the pair items are exactly the same. Moreover, in agreement with the second hypothesis, when choices are made between pairs of items with similar ratings, the true preference for the chosen item is likely to be higher than the true preference for the rejected item. Therefore, when, in the last part of the experiment, subjects are asked to rate the items again, the rating of the chosen item will tend to increase, while the rating of the rejected item will tend to decrease, due to a regression to the mean effect. This argument casts doubt on the existence of choice-induced preference change. Therefore, several control conditions have been suggested. First, Chen and Risen suggested evaluating the amount of information about true preference revealed by the choice by introducing a "rate-rate-choose" condition (Chen and Risen, 2010). In this condition, subjects rate items twice in a row and provide their choices at the end of the experiment. Therefore, changes in preferences between the two ratings cannot be attributed to choices but only to noise in the ratings, as choices are performed after the second rating. This condition has been added to the classic free-choice paradigm in several studies (Izuma et al., 2010; Sharot et al., 2012). Note that the study by Izuma et al (2010) is the only fMRI study controlling for the artefact highlighted by Chen and Risen (2010). Other authors used a blind choice paradigm (Egan et al., 2010; Sharot et al., 2010; Nakamura and Kawabata, 2013). In this task, subjects are asked to choose between items that they do not see. For example, in the study by Sharot et al (2010), participants were asked to choose between two masked vacation destinations, that they could not consciously perceive (actually, no vacation destination was presented and the choice was randomly performed by the computer). This elegant design prevents the choice from revealing any information about true preferences and is thus a nice way to correct for the artefact identified by Chen and Risen. A similar approach has been used by Johansson and colleagues: they used the choice blindness effect to dissociate preferences from choices (Johansson et al., 2013). In this experiment, the feedback subjects received about their choices was sometimes opposite to their actual choices. Interestingly, the observed change in preferences was related to the choices participants believed they made, based on the feedback, but not on their actual

choices. One last strategy, the "implicit-choice paradigm", has been suggested to correct the artefact highlighted by Chen and Risen (2010). In this task, choices are organized in pairs: two similarly rated items (*a* and *b*) are associated with an item rated higher (*h*) and an item rated lower (*l*) respectively. Therefore, the item *a* has a high probability of being rejected, while the item is likely to be chosen. By avoiding the direct comparison, no additional information about preferences for *a* and *b* is revealed. However, these different new paradigms also suffer from methodological problems, as reviewed by Izuma and Murayama (2013) and Salti et al. (2014). Therefore, there is a need for well controlled experimental paradigms.

Another major issue in this literature, which remains unsolved, is what type of cognitive processes underlies choice-induced preference change. There are two opposite theoretical accounts. On the one hand, Festinger's original theory as well as many of its revisions implicitly suggest that cognitive dissonance effects rely on episodic memory (Festinger, 1957; Bem, 1972; Cooper and Fazio, 1984; Steele, 1988; Stone and Cooper, 2001; Heine et al., 2006; Salti et al., 2014). On the other hand, a series of experimental studies challenge this view by showing the existence of choice-induced preference change in amnesic patients (Lieberman et al., 2001) or across very long periods of time (Sharot et al., 2012) or by assessing directly the role of memory (Coppin et al., 2010). These results suggest that cognitive dissonance resolution is an automatic process triggered by the choice, which does not involve high-level processes, such as episodic memory. To dissociate these two opposite theoretical accounts, a crucial aspect would be to elucidate when exactly the change in preference occurs: the first theoretical account would be supported if the change is only observed when subjects are asked to re-rate the items; conversely, the second theoretical account would be supported if the change can be observed immediately after the choice.

In a previous behavioral study (Salti et al., 2014), we used an experimental paradigm controlling for the artefact highlighted by Chen and Risen (2010) and we identified a crucial role of episodic memory for choice-induced preference change. Indeed, we used a design including the "rate-

rate-choose" condition and the difference between the experimental condition and this control condition was significant only when subjects remembered their choices. However, we could not address the issue of the timing of attitude change using these behavioral data. Therefore, we designed the current fMRI study and addressed this issue by capitalizing on previous studies suggesting that brain areas associated with subjective values are not only activated during rating tasks, but also during passive viewing or task that are not related to preferences (Kim et al., 2007; Lebreton et al., 2009; Tusche et al., 2010, 2013; Levy et al., 2011). These brain regions, also known as the "brain valuation system", include the striatum and the ventromedial prefrontal cortex (vmPFC), which correlate with subjects' preferences as measured by ratings or choices (for a review, see Kable and Glimcher, 2009; Bartra et al., 2013). To do so, in the present study, we introduced an incidental task within the free-choice paradigm. Right after the first choice, subjects were presented with the same items as in the ratings and were asked to pay attention to these items in order to detect any repetitions in the stream of stimuli. We hypothesize that, given that the brain valuation system is activated automatically, this task could help us disentangle between the two theoretical accounts of cognitive dissonance: if the activity in the brain valuation system during the incidental task is similar to the one observed during the first rating, it suggests that the choice did not trigger automatically a change in preferences. Conversely, if the activity in the brain valuation system during the incidental task is similar to the one observed during the second rating, it suggests that the choice did trigger a change in preferences, before subjects' second explicit rating.

The second aim of the current study is to refine the role of memory in cognitive dissonance resolution. Our behavioral experiment suggested that choice-induced preference change can only be observed when choices are remembered, as assessed by an explicit memory test at the end of the experiment. However, we could not test whether subjects were actually retrieving their choices when performing the second rating. In the present study, participants were scanned during all the steps of the experiment, including the second rating. We hypothesize that brain areas involved in memory retrieval (for review see Squire, 1982; Eichenbaum et al., 2007;

Preston and Eichenbaum, 2013), such as the hippocampus and medial prefrontal cortex (mPFC), should be activated during the second rating. This result would suggest that participants are using the information that they remember about their choices to perform the second rating.

In the present chapter, I will present preliminary results of this fMRI study, as this work is still ongoing.

3. Materials and Methods

3.1. Subjects

Twenty six participants were recruited through posted advertisements. Six participants were eliminated due to a technical problem with the fMRI machine. The reported analyses were therefore based on 20 subjects (12 males, 8 females; age range 20-30). Participants completed a screening form for significant medical conditions. All participants gave informed consent and were paid 80 euros for their participation.

3.2. Stimuli and procedure

3.2.1. Overview

The experiment was composed of 5 consecutive blocks, during which fMRI data were acquired (Fig. 5.1): two rating blocks (Rating 1 and 2) and two Choice blocks (Choice 1 and 2) and one repetition detection block. Finally, a memory test and a familiarity assessment were performed after the scanning sessions.

3.2.2. Stimuli

Stimuli consisted of 120 colored images of potential vacation destinations. In Rating blocks and in the repetition detection block, one image was presented at the center of the screen with the destination name printed below the image (font size= 30) in each trial. In Choice blocks, two images and their respective names were presented to the left and to the right of the screen, in each trial. The order in which stimuli were presented within each block was random.

3.2.3. Rating 1

Subjects viewed 120 destinations and were asked on each trial to report how much they would like to spend their vacation in that given destination on an eight point scale (1= "I do not want to go there at all", 8= "I would love to go there"). Each trial began with a fixation point that lasted 2 to 3 seconds. Then the destination appeared for 3 second. Subsequently, subjects were presented with an image of two schematic hands with a number from 1 to 8 above each finger (except the thumbs) that randomly varied positions from trial to trial (Fig. 5.1). Subjects were instructed to respond with the finger corresponding to the rating they would like to give. This technique allowed us to avoid different handedness related biases. This screen was presented until response. If subjects answered in less than 3 seconds, a fixation dot was presented, so the interval between the end of the destination presentation and the next trial lasted at least 3 seconds.

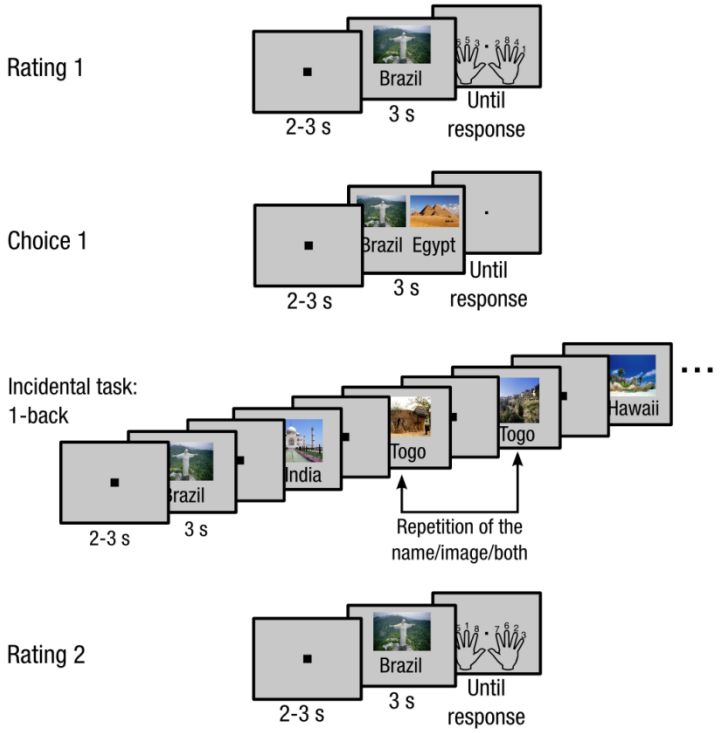

FIGURE 5.1: EXPERIMENTAL DESIGN

The experimental paradigm includes 6 steps. First, subjects are asked to rate all items. Then they choose between similarly rated items. The third step is an incidental task: subjects are asked to pay attention to the items that are presented and to detect any contiguous repetition of the name, the image or both. Then they are asked to rate all items again and to choose between the second half of the images. Finally, they are presented with a memory test, which assess both their objective and subjective memory.

3.2.4. Choice 1

Following the Rating1 block, subjects were presented with pairs of destinations they had rated equally and had to indicate with a button press at which one they would rather take their vacation. Note that in each of the Choice tasks, subjects viewed 25 pairs of destinations based on ratings given in Rating 1. This means a total of 100 destinations out of the 120 were used to form the Choice blocks. The 20 remaining destinations (the same ones for all subjects) were used as repetition targets for the following repetition detection task. Each trial began with a fixation point that lasted 2 to 3 seconds. Then the paired destinations appeared for 3 seconds, followed by a fixation point, during which subjects had to give their choice. This fixation screen lasted at least 3 seconds and at most until subject's response.

3.2.5. Incidental task: Repetition detection task

In this task subjects were presented with all 120 destinations as each appeared at the screen with the same presentation time as rating 1 and rating 2. 1/6th of the time either the destination name or the image, or both were repeated contiguously (only for the 20 images mentioned in Choice 1 description). Subjects had to press a button to signal that they detected this repetition. This repetition detection task was implemented in order to make sure subjects were attentively attending all destinations and their respective names. The main aim of this block was to induce automatic valuation of the different destinations.

3.2.6. Rating 2

This block was strictly identical to Rating 1.

3.2.7. Choice 2

This block was strictly identical to Choice 1, but included the other half of the stimuli.

3.2.8. Postscanning questions

After the scanning sessions, participants were asked to perform two additional tests outside the fMRI. First they performed a memory test concerning the choices they had made in Choice 1 and Choice 2 sessions. In order to avoid any explicit memorization of the items, subjects were not informed about this memory test at the beginning of the experiment. Every trial began with the

appearance of a destination and its name at the center of the screen. Below the item, were stated the options "chosen" and "rejected" at the left and right of the screen. Using the keyboard arrows, subjects had to indicate whether they remember choosing or rejecting that given item during the choice tasks. This question tested objective memory of the choices. After giving their answer, appeared a sentence at the center of the screen "Are you sure of your response?" underneath which were presented the two options " I am sure" and "I guessed" at left and right of the screen. Again subjects had to use the keyboard arrow to specify their answer. This second question tested the more subjective aspect of subjects' memory. There was no time limitation and the trials concerned all 100 items that were seen during Choice 1 and Choice 2.

Finally, subjects rated the familiarity of each of the 120 destinations using an eight-point scale (1= "I do not know this country at all", 8 = "I am very familiar with this country").

3.3. Behavioral analysis

Subjects' responses during both ratings were extracted using Matlab 2013a (Mathworks). A linear mixed model was used to analyze the relationship between memory and spread in this task. Indeed, as opposed to a classical ANOVA, such a model does not require prior averaging of the spread for each subject and thus offers the possibility to handle the heteroskedasticity related to the unbalanced number of items in each condition. Significance of the fixed effects was assessed using the Kenward-Roger approximation for degrees of freedom of the denominator, with the 'lmerTest' package in R. Restricted analyses were performed using t-tests.

3.4. Functional MRI Data Acquisition and Analysis

Functional imaging was conducted using a 3 Tesla Siemens Verio scanner and a multiband echo-planar imaging sequence sensitive to brain oxygen-level-dependent (BOLD) contrast to produce 45 continuous 2.5-mm thick transaxial slices covering nearly the entire cerebrum (repetition time = 1.022 ms; echo time = 25 ms; flip angle = 60°; field of view = 100 mm2 view = 100 mm2 80 × 80 matrix; voxel dimensions = 2.5 × 2.5 ×2.5 mm). A high-resolution anatomical T1-weighted image was also acquired for each subject for anatomical localization.

fMRI data were analyzed using SPM8 (SPM8, Wellcome Department of Cognitive Neurology, London, United Kingdom) toolbox. Functional images were realigned, unwarped using the FSL "Topup" toolbox in order to correct EPI distortions due to B0 field inhomogeneity (Andersson et al., 2003) and then normalized into Montreal Neurological Institute standard stereotactic space. The normalized fMRI data were spatially smoothed with a Gaussian kernel of 8 mm (full-width at half-maximum) in the x, y, and z axes.

Two main analyses were performed. GLM 1 aimed at identifying brain regions that were positively correlated with subjects' reported preference for each destination. This model was estimated using data from Rating 1 and Rating 2 and included 3 regressors: (i) each vacation destination stimulus onset, (ii) stimulus onset modulated by subjects' reported preference for each destination and (iii) stimulus onset modulated by reported familiarity for each destination. GLM 2 was designed to examine the neural correlates of our behavioral finding showing a significant spread difference between the RCR and RRC conditions only for memorized items. To this aim, we compared brain activity for memorized versus forgotten items in the RCR compared to RRC conditions. Trials from Rating 2 were classified into 4 conditions: (i) Remembered items in the RCR condition, (ii) Forgotten items in the RCR condition, (iii) Remembered item in the RRC condition, (iv) Forgotten items in the RRC condition. The anatomical region of interest (ROI), including bilateral hippocampi and bilateral parahippocampal gyri, was defined using the SPM WFU PickAtlas tool (Maldjian et al., 2003, 2004).

4. Results

4.1. Behavioral results

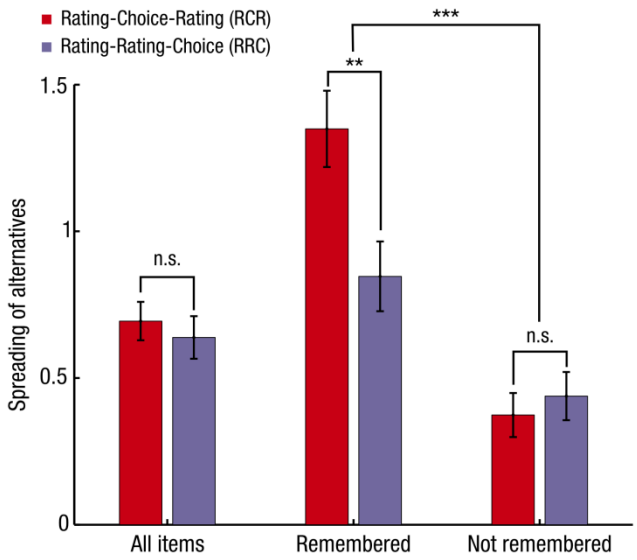

FIGURE 5.2: BEHAVIORAL RESULTS

The difference between the condition of interest, RCR, and the control condition, RRC, was significant only for remembered items.

First, our behavioral data revealed that the spreading of alternatives for the condition of interest, "rate-choose-rate" (RCR), was not significantly different than that of the control condition "rate-rate-choose" (RRC) ($t(19) = 0.5836$, $p = 0.5664$ – Fig. 5.2). We tested the effect of memory using a linear mixed-effects model including the spread for each pair of items for each subject, with the 'Condition (RCR/RRC) and the 'Memory' (remembered/forgotten) factors defined as fixed factors and subjects defined as a random factor. A main effect of memory was found ($F(1,943) = 51.87$; $p<10\text{-}11$), as well as a main effect of condition($F(1,980) = 6.71$; $p=0.009$). Interestingly, the critical interaction between memory and condition was significant ($F(1,992) =11.14$, $p<0.001$ – Fig. 5.2). Restricted analyses revealed that in the RCR condition, the difference in spread for remembered versus forgotten items was significant ($t(19) = 6.6521$, $p =$ 10-5). Surprisingly we also observed a difference in spread for remembered compared to forgotten item from the RRC condition $t(19) = 2.9717$, $p = 0.008$. This result though unexpected is consistent with our previously obtained result in a similar experiment (Salti et al. 2014) emphasizing yet again the role of attention in this mechanism. Importantly, the difference in spread between RCR and RRC was only significant for remembered items ($t(19) = 3.1143$, $p =$ 0.005), but not for forgotten items ($t(19) = -0.5657$, $p = 0.5782$)(Fig.2). This result robustly

replicates our previous findings and confirms the role of memory as a mandatory condition to observe choice induced preference change.

5. Imaging results

We first tried to identify brain regions that encode the subjective preference for the various destinations. Using subjects' rating values as a covariate, we conducted a parametric modulation analysis on data from all 240 trial from Rating 1 and Rating2. Statistical threshold was set at $p < 0.001$ for height (uncorrected) and $p < 0.05$ Family wise error (FWE) cluster correction. Brain areas showing significant positive correlations with preference ratings included the left ventromedial prefrontal cortex and right ventral striatum (Fig. 5.3a).

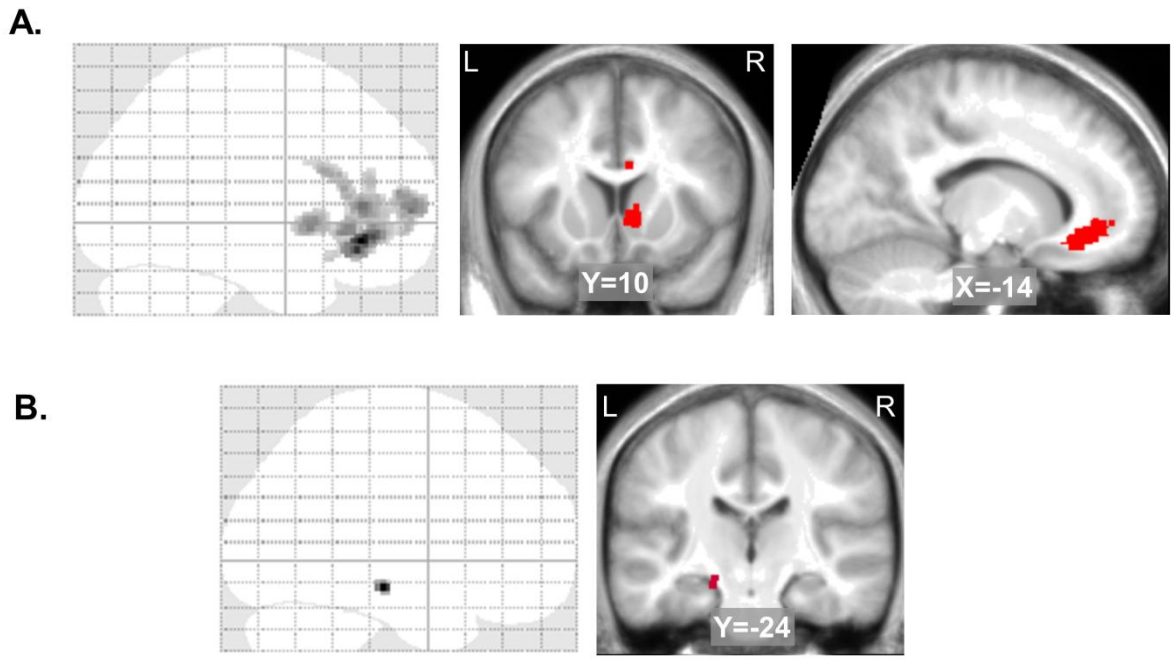

FIGURE 5.3: IMAGING RESULTS

A. *Brain activity related to preference for vacation destinations. Regions in which BOLD response was positively correlated with participants' preference ratings ($p < 0.001$ for height (uncorrected) and $p < 0.05$ Family wise error (FWE) cluster correction) included the right striatum and the left ventromedial prefrontal cortex.*
B. *Left hippocampus activation for the remembered compared to the forgotten items of the RCR compared to the RRC condition during Rating 2((ROI FWE-corrected $p < 0.05$).*

Next, we tested whether the memory effects observed at the behavioral level were supported by brain activations. To this aim, we contrasted brain activity in memory-related areas (ROI including bilateral hippocampi and bilateral parahippocampal gyri) between remembered and forgotten items, in the RCR compared to the RRC conditions. This interaction was significant in the left hippocampus (ROI FWE-corrected $p < 0.05$ – Fig. 5.3b).

6. Discussion

We studied the neural mechanisms underlying choice-induced preference change in the free-choice paradigm.

We used a control condition, evaluating the statistical artefact identified by Chen and Risen (2010) and found that attitude change was only observed when participants remembered their choices. Interestingly, this result confirms the findings of our previous study (Salti et al., 2014) suggesting a crucial role of memory in choice-induced preference change. Moreover, we also replicated the significant difference between remembered and forgotten items in the RRC condition. As suggested in our previous study, this unexpected result could reflect a better attentional engagement during to-be-remembered choices, which thus reveal more accurate information about subjects' preferences, in agreement with Chen and Risen's framework. Importantly, note that in the present study, when considering all the items together, the difference between RCR and RRC was not significant, contrary to what was observed in our previous study. This difference can be explained by the design of both experiments: the current paradigm includes a repetition detection task, right after the first choice, so the experiment lasts longer than our previous one and the first choice is further away in time from the memory test. These differences can create a decrease in memory performance, which could explain why RCR and RRC conditions are not significantly different when considering all items together. Indeed, the difference between these conditions is the combination of the differences between RCR and RRC for remembered and forgotten items. When a lot of pairs are correctly remembered, the average difference across all pairs of items will tend to be higher than when memory performance is low. This result highlights the fundamental role of memory in this experiment. These analyses were performed using objective memory, but it would be interesting to explore this issue using also subjective memory measurements that we also acquired, in order to understand if false memories also trigger choice-induced preference change. If it turns out to be true, this result would reinforce the claim concerning the role of memory, as it would suggests

that it is not the choice in itself which triggers attitude change, but only what participants remember choosing.

Preliminary analyses of fMRI data obtained in this paradigm revealed that activity in the right ventral striatum and the left vmPFC was correlated with subjects' explicit preferences, as measured by the two ratings. These brain regions have been repeatedly identified as the "brain valuation system" (Kable and Glimcher, 2009; Lebreton et al., 2009; Bartra et al., 2013). Some of the previous studies exploring the neural mechanisms of cognitive dissonance identified only the ventral striatum as correlated with subjects' preferences, but did not find an implication of the vmPFC (Sharot et al., 2009; Izuma et al., 2010). This difference between our results and previous studies could be explained by differences in the type of stimuli that were used and in instructions. Izuma et al (2010) used food items as stimuli and asked participants how much they like them, while Sharot et al (2009) used vacation destinations and asked participants to imagine how happy they would be if they were to vacation at that location. The difference between the striatum and the vmPFC in the valuation process is still debated, but it has been suggested that the striatum would encode action-based value responses, while the vmPFC would encode more abstract good-based value responses (Kable and Glimcher, 2009). We can speculate that our instructions might trigger both types of encoding, while this was not the case in previous studies which were more oriented towards actions.

Our last preliminary result is that the hippocampus, known to be involved in memory processes (Squire, 1982; Eichenbaum et al., 2007; Preston and Eichenbaum, 2013), is activated during the second rating. Interestingly, this activation is higher in the RCR than in the RRC condition. This result can be interpreted as evidence for the existence of memory retrieval processes during the second rating. This interpretation strengthens our previous interpretation of the behavioral results (Salti et al., 2014). Indeed, the difference in spreading of alternatives observed between remembered and forgotten pairs of items does not provide direct evidence that subjects were actively retrieving information about their choices during the second rating. Therefore, our

behavioral results are only suggestive of the implication of memory in choice-induced preference change and do not allow us to dissociate between the two alternative theoretical accounts presented in the introduction. In this context, the activation of the hippocampus during the second rating constitutes another important evidence of the role of memory. However, it remains to be established in future analyses that this activation is specific to the second rating and does not occur in previous stages, such as the incidental task, in order to provide a plausible mechanism of choice-induced preference change, based on the retrieval of choice information during the second rating.

Several issues are not yet addressed by the preliminary results presented here. First, we have identified brain areas in which the activity correlated with participants' ratings, but we have not explored yet which brain regions reflect the choice-induced preference change observed for remembered pairs of items. Based on the only previous study exploring neural correlates of choice-induced preference change in a controlled experimental paradigm (Izuma et al., 2010), we expect that the striatum and the vmPFC should be correlated with the spreading of alternatives during the second rating. Note, however, that, as highlighted in our previous study (Salti et al., 2014), in Izuma et al's experiment, subjects were presented with a reminder of their choice during the second rating, so this might have influenced the result related to the correlation with the spread during this rating. Second, we have not analyzed yet brain activity during the incidental task. This task was introduced to provide a better understanding of the timing of choice-induced preference change. We will test whether preference change can already be detected during this incidental task or if it only appears during the second rating. The first hypothesis would support an automatic account of choice-induced preference change, while the second combined with our previous behavioral and fMRI results would support the involvement of high-level cognitive processes, such as memory.

The preliminary results obtained so far in this fMRI study replicate our previous behavioral results and provide new evidence for the role of memory in choice-induced preference change.

However, they do not yet allow us to dissociate between the two possible mechanisms underlying this effect. We will pursue the analysis of these data, notably by exploring brain activity during the incidental task, to better understand the neural underpinning of this cognitive dissonance effect.

Supplementary data

Location	MNI coordinates of the peak			Mean T	Cluster size
	x	y	z		
Left vmPFC	-14	34	-10	4.05	152
Right striatum	10	10	-4	4.1	72
Medial anterior PFC	-18	58	4	4.09	223

TABLE S.1: AREAS CORRELATED WITH SELF-REPORTED PREFERENCE DURING RATING 1 AND RATING 2

Statistical threshold was set at p < 0.001 for height (uncorrected) and p < 0.05 Family wise error (FWE) cluster correction.
BA, Brodmann area; vmPFC, Ventro-medial prefrontal cortex; PFC, prefrontal cortex.

CHAPTER 6. GENERAL DISCUSSION

1. Summary of the results obtained during this PhD

The aim of this work was to better understand how subjects detect inconsistencies (both in the environment and in their own behavior) and adapt their behavior accordingly, but also how this process interacts with conscious access. We performed four experimental studies to address these issues.

1.1. Local and global auditory novelty processing

In the first experimental study presented in this thesis (Chapter 2), we aimed at understanding how subjects process inconsistent information in the environment. To do so, an experimental paradigm combining two levels of auditory regularities was used and the detailed dynamics of novelty detection in these two levels was studied using intracranial EEG recordings.

Combining event-related potentials (ERPs), time-frequency and functional connectivity analyses, we found that local novelty detection was associated with an early process confined to the temporal lobe, while global novelty detection involved late cognitive processes relying on the coordinated activity of distributed brain regions.

Signatures of global novelty detection were strikingly similar to the results observed when analyzing conscious processing of visual unmasked stimuli (Gaillard et al., 2009). This parallel across different modalities suggests the existence of common mechanisms of conscious access (Sergent and Naccache, 2012). Moreover, these results are in agreement with theories of consciousness proposing that conscious perception is related to the integration of information across long distances (Dehaene and Naccache, 2001; Dehaene et al., 2006; Dehaene and Changeux, 2011). Note however that in this first study patients did not provide subjective reports of their conscious access to the global regularity. The proposed link between the detection of the global regularity and consciousness is based on previous studies using the same experimental paradigm (Bekinschtein et al., 2009; Faugeras et al., 2011, 2012; Wacongne et al.,

2011). These studies report that signatures of global novelty detection are not present when subjects cannot report the global regularity, as opposed to signatures of local novelty detection which are present independently of subjects' attention. In this study, it is also interesting to notice that stimuli are always consciously perceived, but what determines the presence of markers of global novelty detection is conscious access to the regularity itself.

Local and global novelty detections highlight the existence of two mechanisms allowing subjects to process information contracting their priors. Indeed, once the regularity has been learned and the prior probability of the standard stimulus has become very high, the mechanisms underlying deviant stimuli processing are different if they violate the local or the global regularity. The interaction between these levels of regularity and consciousness gives rise to a second question: can an inconsistency which is not perceived consciously influence behavior?

1.2. Acquisition and transfer of strategies based on conscious and non-conscious stimuli

This issue was addressed in our second study (Chapter 3). Using an experimental paradigm derived from the Stroop task, we varied the frequency of incongruent trials, in order to study cognitive control, and more precisely to test adaptation to conflict in behavior and EEG measures. Importantly, we manipulated the visibility of the cue, so in certain trials, subjects did not consciously perceive the presence of a conflict between the cue and the target.

We found three interesting results. First, we replicated results of Merikle and colleagues, who found that adaptation to conflict was only present when the conflict was perceived consciously (Merikle and Cheesman, 1987; Merikle et al., 1995; Merikle and Joordens, 1997). Second, we hypothesized that the acquisition of the new strategy (i.e. adaptation to conflict) may not be possible when the conflict remains non-conscious, but the new strategy could be transferred from conscious trials to non-conscious trials. We found evidence of such transfer only in the modulation of the P300 in masked trials included in mostly unmasked blocks. Note that no evidence for strategy transfer was found in reaction times. Third, adaptation to conflict in mostly incongruent blocks was modulated from trial to trial, suggesting that the strategy was evaluated

based on recent history. These changes in subjective strategy adherence were only observed in unmasked trials and might thus require conscious access to the stimuli.

It would be interesting to study this last point in more detail and to understand whether subjects need to be conscious of the stimuli themselves or whether they need to consciously experience a conflict, as suggested by Desender et al. (2013), even when stimuli remain unconscious. Moreover, to better understand the dynamics of subjective strategy adherence, it would be relevant to assess subjects' confidence in the strategy based on adaptation to conflict on each trial. Note that in this study, subjects were explicitly instructed about the existence of a high proportion of incongruent trials but their behavior suggests that their reliance on this instruction fluctuates. Evaluating subjects' confidence on each trial would allow testing a potential dissociation between stimuli reportability and strategy reportability.

These results highlight how subjects can adapt their behavior in light of regularities in their environment. When this adaptation involves cognitive control functions, it seems to be limited to conscious stimuli, although this point is highly debated (Desender and Van den Bussche, 2012; Kunde et al., 2012; van Gaal et al., 2012). Note that in the study presented here, the task we used required access to the semantic features of the cue and the use of this information to adapt behavior within the very short time separating the cue from the target. The difficulty of this task might have influenced our results, as studies showing conflict adaptation to unconscious stimuli are based on paradigms using simpler stimuli, e.g. arrows. Finally, in future studies, it would be interesting to study not only the reportability of stimuli, but also the reportability of the strategic adaptation, to provide a better understanding of this process.

1.3. The crucial role of explicit memory in cognitive dissonance

The first two experimental studies allowed us to test how subjects detect and use patterns in their environment, but we could not assess their internal consistency. To do so, we used the framework of cognitive dissonance in the third (Chapter 4) and fourth (Chapter 5) studies and

investigated how subjects deal with inconsistencies in their own behavior by updating their attitudes.

Using the free-choice paradigm, we found evidence for the existence of choice-induced preference change after correcting for different confounds that were previously reported (Chen and Risen, 2010) or that we identified in the literature. Interestingly, choice-induced preference change was actually only present when subjects remembered their choices (Chapter 4). This effect of memory was replicated in three experiments. This result contradicts previous studies which were affected by the statistical artefact identified by Chen and Risen (Lieberman et al., 2001; Coppin et al., 2010; Sharot et al., 2012). The involvement of memory of the choice in the cognitive dissonance effect suggests that subjects need to be conscious of the conflict between their preferences and their choice to adapt their behavior to this inconsistency. Interestingly, this result is in agreement with the original formulation of cognitive dissonance theory (Festinger, 1957), but not with new empirical data suggesting that choice-induced preference change is an automatic process triggered by the choice. Note that this paradigm did not allow us to test other theories of cognitive dissonance, e.g. we could not test the implication of the self.

This result was based on a memory test performed at the very end of the experiment, but we could not test whether choices were actually recalled during the second rating. Moreover, memory was assessed using a forced-choice between "chosen" and "rejected", so no information about subjects' confidence in their memory was provided. Therefore, we conducted the fourth study presented in this thesis (Chapter 5).

Using fMRI, we tested the activation of memory related brain areas to assess if subjects tried to recall their choices during the second rating. Moreover, subjects were presented with both an objective and a subjective memory tests. Finally, to refine our understanding of choice-induced preference change, we introduced a passive viewing task between the choice and the second rating to test if preferences changed immediately after the choice or if this change was related to the re-assessment of preferences. Although the analysis of these data is still ongoing, we

identified brain areas correlated with subjective values, as measured by the ratings, and we found that the memory network is activated during the second rating. We now need to further analyze the data to identify correlates of choice-induced preference change and to test when this change occurs. We also acquired intracranial EEG data using the exact same paradigm, in order to refine our understanding of the fine dynamics of the process and their analysis is currently ongoing.

In the studies presented in this manuscript, we showed that memory of the choice was important to observe choice-induced preference change. One interesting issue is whether memory is necessary and sufficient for choice-induced preference change. In collaboration with Mariam Chammat and Sébastien Allali, we are currently trying to address this issue by testing different groups of patients. First, amnesic patients suffering from mild cognitive impairment (MCI) were tested using the paradigm presented in Chapter 4 but including a subjective memory assessment. So far, 17 patients were included and their results confirmed the role of memory. The pattern was fairly similar to what we reported in control subjects, although memory performance was lower (mean memory rate = 30%, comparison with control subjects included in the MRI experiment: $t(35) = 2.14$, $p = 0.03$): there was a significant difference between RCR and RRC for remembered items, but not for forgotten items (Fig. 6.1). Note that, when comparing RCR to RRC independently of the memory, there was no significant difference. Second, the hypothesis that memory is actually not sufficient for choice-induced preference change was tested in two other groups of patients: patients suffering from fronto-temporal dementia (FTD) and schizophrenic patients. We chose these two groups of patients because we think that on top of memory, preserved executive functions and internal consistency are needed to observe choice-induced preference change. The data from these patients are currently being collected and analyzed and we will be able to test this hypothesis.

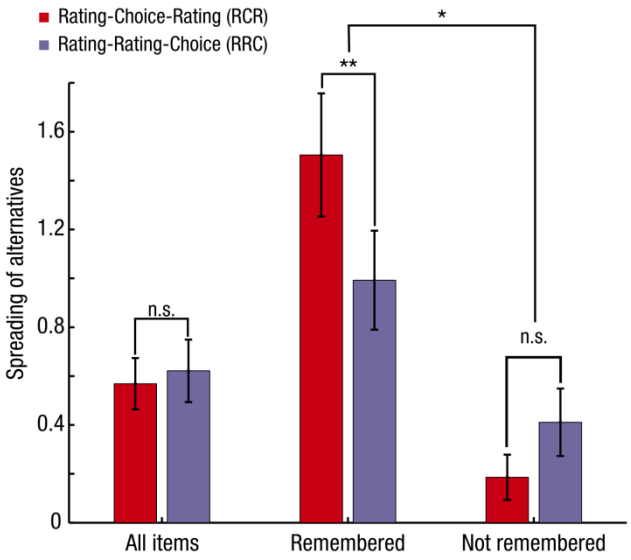

FIGURE 6.1: BEHAVIORAL RESULTS

The difference between the condition of interest, RCR, and the control condition, RRC, was significant only for remembered items.

In sum, we were interested in cognitive dissonance as a way to test how subjects update their interpretations when confronted with inconsistencies in their behavior. Given the main result of our experiments, namely the role of explicit memory in choice-induced preference change, one can wonder whether the free-choice paradigm is actually a good model to address this issue. Indeed, further analysis of our fMRI data should reveal whether choice-induced preference change occurs right after the choice or if it occurs because subjects are asked to re-rate the items. In the second hypothesis, the change in subjects' reported preferences would not necessarily reflect a change in their real values, but could just reflect an adaptation to task demands without any update of values.

1.4. Integration of all the results

The results presented in these four studies can be summarized in the following diagram (Fig. 6.2). When facing inconsistencies in their environment or in their own behavior, subjects first have to identify them. The first study provided evidence for the existence of two different detection mechanisms, an automatic one and a more strategic one. Then, subjects can adapt their behavior to the inconsistencies they detected, as suggested by the last three studies.

Interestingly, in these studies, no evidence of behavioral adaptation was found when the inconsistencies remain non-conscious.

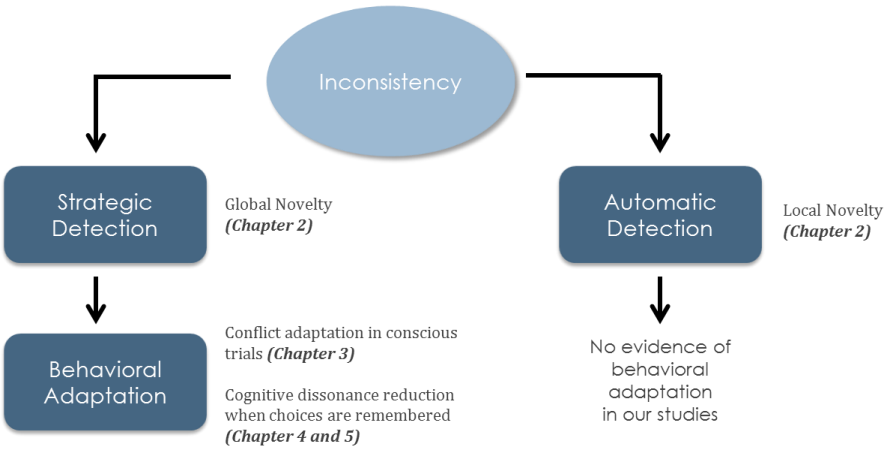

FIGURE 6.2: DIAGRAM SUMMARIZING THE RESULTS

2. General limits

In this thesis, I have presented four studies that try to address how we identify inconsistencies in our environment and in our own behavior and how we update our models of the world in response to them. These studies approach these questions using very different paradigms and the relevance of their integration within one project should be discussed.

The first and the second studies rely on objective markers, such as reaction times or EEG signatures, indicating that subjects processed the regularities experimentally introduced in the environment. The existence of these regularities in stimuli was explicitly stated in the instructions for both studies. Conversely the third and the fourth studies rely on participants' subjective reports about their preferences. Moreover, subjects are not explicitly told to identify inconsistencies in their own behavior, rather they are asked to perform rating and choice tasks which could be unrelated. For example, participants are not informed that their first rating is used to determine which choices they are going to perform during the second step of the experiment. These differences between the two series of experiments are interesting because they allow testing detection of inconsistencies and behavioral adaptation in different contexts, but they also limit the validity of the comparison of the results obtained. For example,

adaptation to inconsistencies in the environment or subjects' own behavior was studied in this thesis. However, the mechanisms of cognitive control modulation identified in the second study cannot be easily compared with cognitive dissonance results, given the differences between the two experimental settings.

Moreover, I started the introduction of this thesis by presenting examples of complex interpretations produced by neurological patients and control subjects. It is interesting to question to what extent the four experimental studies I conducted help understanding the formation and the dynamics of these interpretations. First, interpretations were defined as subjective reports reflecting the integration of priors with data sampled in the environment. In the studies presented in this thesis, subjective reports were assessed only sparsely, because they are difficult to quantify. But exploring the narrative aspect of interpretations is a very interesting avenue for future research. Second, as we were interested in the neural mechanisms underlying interpretations dynamics, we chose to use experimental paradigms in which inconsistencies could be easily produced. This choice of strictly controlled task designs reduced the complexity of the interpretations that could be studied. They were indeed restricted to regularities in auditory and visual stimuli and preferences for vacation destinations. Moreover, this approach limited our ability to study in details the subjective dimension of interpretations. Finally, one characteristic of interpretations is that they are coherent and meaningful in the context in which they are formulated. The studies presented in this thesis aimed at understanding what processes allow the identification of inconsistencies and their correction. An interesting perspective for future studies is to also explore why such a consistency is observed.

Two important questions remain in light of the results presented of these four studies and they are related to the two concepts that appear all along this thesis: consciousness and consistency. The first one is what do we need to be conscious of when facing inconsistencies. The second is what can explain our need for consistency. I will present several thoughts about these issues in the following sections.

3. Two major issues raised by this work

3.1. What do we need to be conscious of?

All of our four studies addressed the relation between inconsistencies and consciousness. In the first study, we identified two types of inconsistency detection processes which are differently related to consciousness: the local one is automatic and can be observed even in states of altered consciousness, while the global one is more strategic and is absent when subjects are not able to report the regularity in the stimuli. In the last three studies, we examined adaption to inconsistencies, both in the environment and in subjects' own behavior. One very interesting similarity in these studies is that adaptation was only observed when participants were conscious of the inconsistencies. In the second study, blockwise and trial-to-trial conflict adaptations were only found when the cue was clearly visible. In cognitive dissonance studies, choice-induced preference change was only observed when subjects remembered their choices, suggesting that it occurs only when they could access the inconsistency in their behavior.

Adaptation to inconsistencies both in the environment and in our own behavior thus required conscious access to these inconsistencies, but an interesting aspect that was not directly addressed in our studies is that the process of adaptation to these inconsistencies itself is probably not accessible to consciousness. The most convincing example is cognitive dissonance reduction. In the debriefing following the experiments, it was evident, although we did not quantify it, that subjects were not aware of the change in their preferences. In a review on subjective reports, Nisbett and Wilson (1977) took examples from the cognitive dissonance literature and provided clear evidence that subjects were not conscious of the changes in their attitudes and sometimes they were not even conscious of the negative arousal related to inconsistencies in their behavior. Similarly, Ghinescu et al. (2010) demonstrated that the selection of a particular conflict adaptation strategy could be determined without conscious awareness. They used an Eriksen flanker task (Eriksen and Eriksen, 1974), with four possible stimuli: HHHHH, SSSSS, HHSHH, SSHSS. In this task, participants are asked to indicate the identity of the central letter (either S or H). It has been consistently shown that they are faster

when responding to the first two stimuli in which the distractors point to the same response as the target (congruent stimuli), than to the last two, in which the distractors point to the opposite response relative to the target (incongruent stimuli). This effect is known as the flanker effect. In Ghinescu and colleagues' study, subjects were presented with a cue (the letter A, B or C), which was followed by one of the four stimuli presented above. These cues had predictive values: A indicated with a probability of 80% that the upcoming stimulus would be incongruent, C indicated with a probability of 80% that the upcoming stimulus would be congruent and B indicated that congruent and incongruent stimuli were equally likely. Importantly, subjects were divided into three groups and received explicit, partially explicit or implicit instructions about the validity of the cues. The results showed that participants were able to adapt their conflict adaptation strategy based on the cue, even when they did not receive explicit instructions. Interestingly, using two questionnaires, the authors were able to show that subjects in the partially explicit and implicit instructions groups could not report the contingencies between cues and trial congruence and thus did not explicitly use the cues to select a particular strategy. Note that in this experiment, the cues were clearly visible. This study highlights that similarly to what is found in the cognitive dissonance literature, conflict adaptation strategy can be determined implicitly. Finally, a case study of a posterior split-brain patient conducted in our lab also suggests that strategies are not necessarily consciously initiated by subjects (Naccache et al., 2014). In this study, the patient was asked to perform a matching-to-sample task with Arabic digits: two numerical targets (T1 and T2) were presented, one at a time, and the patient had to indicate if T1 and T2 were identical or different. Note that T1 and T2 could either be presented in the same hemifield (intra-hemifield trials) or in different hemifields (inter-hemifield trials). Moreover, the stimulus onset asynchrony (SOA) between T1 and T2 was manipulated to create two conditions: long and short SOA. The patient performed correctly on intra-hemifield trials but was impaired in inter-hemifield trials. Interestingly, this impairment decreased across blocks in the long SOA condition, but not in the short SOA condition. The patient actually developed an efficient strategy to compare targets presented in different hemifields in the long

SOA condition, by transferring information across hemispheres through his preserved anterior portion of the corpus callosum, as suggested by EEG results. Interestingly, the patient could not report the development of this strategy at the end of the experiment. This study underlines the existence of strategies which are developed automatically on consciously visible stimuli.

Taken together, these results suggest a dissociation between stimulus and strategy awareness. Processes such as cognitive dissonance reduction or adaptation to conflict frequency in the environment seem to require awareness of the inconsistency. Note that this does not necessarily mean that subjects have to consciously perceive all the stimuli. It rather implies that they should experience the conflict, either by clearly perceived the stimuli or by identifying slowing in their reaction times or an increase in their error rates, as suggested by metacognitive accounts of conflict adaptation when cues remain non-conscious (Jáskowski et al., 2003; Kinoshita et al., 2008; Desender et al., 2013, 2014). But importantly, these processes of cognitive dissonance reduction or adaptation to conflict do not seem to be consciously accessible for subjects. It is interesting to notice here the difference between these processes and other high-level cognitive processes such as calculus or problem-solving for example. It has been shown that subjects cannot report the different steps of the latter processes. For example, in a classic study, Maier (1931) asked subjects to tie the ends of two strings hanging from the ceiling of the laboratory. This problem had four possible solutions and some of them involved the use of other objects available in the room. One of the solutions was particularly difficult to find and Maier provided subjects with a cue to help them identify this solution. When asked how they solved the problem, participants never mentioned how this cue helped them, but rather constructed complex interpretations of their thought process. Other examples of such lack of consciousness of high-level cognitive processes steps can be found in the review of Nisbett and Wilson (1977) on subjective reports. The difference between such processes and adaptation to inconsistencies in both the environment and subjects' own behavior is that when solving a problem, participants consciously intend to do so and thus they consciously trigger the process. On the contrary, when confronted with inconsistencies, strategies for dealing with them seem to be triggered implicitly,

without subjects' intention. Therefore, confirming this dissociation between the requirement of conflict awareness and the lack of awareness of the strategy itself, which was not directly addressed in the studies presented in this thesis, constitutes an interesting direction for future studies.

If this dissociation is confirmed, it would define an interesting category of processes, which can be integrated in the taxonomy proposed by Dehaene et al. (2006). In this article, the authors identified four types of processing: unattended subliminal, attended subliminal, preconscious and conscious. In agreement with the Global workspace theory (Dehaene et al., 1998; Dehaene and Naccache, 2001), conscious processing, defined by subjective reportability, was characterized by activation spreading across a wide variety of brain areas, including frontal and parietal ones, and maintained through long distance interactions. The three other types of processing could not lead to subjective reports. Unattended subliminal processing was characterized by very weak brain activation and little influence on behavior via priming. Attended subliminal processing was defined by a strong feedforward activation, which could influence behavior as indexed by priming, but which decreased with depth of processing. Finally, preconscious processing was characterized by intense activation in sensori-motor processors and local synchrony, without broadcasting to more widespread brain areas, in particular frontal and parietal cortices. All these categories of processing relate to perception. Adaptation to inconsistencies is related to higher order cognitive functions and could constitute a new category in this taxonomy. This category would include processes, which require conscious perception of the stimuli they rely on, but which are not consciously accessed themselves. However, further investigation would be needed to understand to which of the three categories of non-conscious processing they belong.

3.2. What can explain our need for consistency?

The second concept that was explored in all four studies is consistency, both in the environment and in subjects' behavior. In this section, I will discuss the different timescales on which we observed consistency across the four experiments. Then, I will try to identify what differentiates

consistency in the environment and in subjects' behavior. Finally, I will discuss why we seek consistency in the environment, but also more speculatively in our behavior.

3.2.1. Different timescales of consistency

Across the four studies reported in this thesis, we could identify detection of inconsistency on different timescales. The local novelty detection was associated with regularity on a short timescale that could be identified by the comparison of the current stimulus to the previous one. This process indexed by the mismatch negativity (MMN) is thought to involve echoic memory, which spans only over a few seconds (Darwin et al., 1972; Winkler et al., 1993; Näätänen et al., 2001, 2007). It is interesting to notice that when the stimulus onset asynchrony (SOA) in an auditory stream increases too much, the MMN in response to deviant stimuli seems to disappear. This effect is actually due to the presence of an equally strong novelty response in all trials, both standard and deviant (Pegado et al., 2010), preventing the detection of the MMN which is computed by the subtraction of ERP responses to deviant and standard stimuli. In this case, every time a new stimulus is presented, the echoic memory buffer is already empty so it is always perceived as new. The second timescale that was relevant in our studies was spanning over tens of seconds to one or two minutes. This timescale was the length of a block of trials. To detect global novelty, subjects had to integrate the different series of sounds across trials. A habituation stage, composed only of global standard trials, was first presented and lasted about 30 seconds. This period allowed participants to identify the global pattern in the auditory stream. Similarly, in the second study, subjects had to integrate conflict information across different trials to extract conflict frequency. Note however that participants were told that incongruent trials would represent 80% of the trials. But as revealed by the trial-to-trial dynamics of conflict adaptation, the subjective adherence to this instruction seems to change over time, depending on the characteristics of the previous trial. This effect suggests that participants evaluated conflict frequency based on their own experience and thus integrated information across several trials. In these two examples, this integration spans over several tens of seconds and involves working memory to detect stimuli which are inconsistent with the

identified regularity. Finally, in the cognitive dissonance studies, we found that memory of the choice was important to observe choice-induced preference change. This result highlights that subjects only adapted their behavior to inconsistencies between their preferences and their actions, when they remembered them. This supposed that they were able to maintain the information about their previous behavior during several minutes, from the choice block to the second rating block, so approximately during five to ten minutes. Note that the memory test was performed at the very end of the experiment and assessed the memory for choices which occurred up to fifteen minutes before. In these studies, short-term episodic memory was thus involved to allow subjects to detect inconsistencies in their own behavior and adapt their responses accordingly. We can speculate, in agreement with Festinger's original formulation of the cognitive dissonance theory (1957), that the identification of inconsistencies based on this short-term episodic memory triggered a negative arousal state that would in turn elicit preference change.

These different results highlight that we seek consistency at several levels, which can be dissociated by the timescales on which they operate. They also underline the implication of different types of memory.

3.2.2. Consistency in the environment vs. internal consistency

In the four experiments presented in this thesis, we chose to study both how subjects seek consistency in their environment and in their own behavior. An interesting point of discussion is to understand how these two types of consistencies might differ.

In their environment, participants seek coherent patterns, even when the sequence of stimuli is perfectly random, as seen in the introduction (Yellott, 1969). Thus, this identification of patterns is a way to manage inconsistencies and uncertainty in the environment by updating priors in a given environment. Indeed, when facing inputs which are inconsistent with their models of the world, subjects try to integrate them in a new model. A typical example is provided by the second study we conducted. In this experiment, subjects were presented with congruent and

incongruent trials. The default response associated with a given cue is to prepare the action it points to. This is based on the default model of the world subjects hold. But in this experiment, incongruent trials were very numerous. Subjects were thus frequently confronted with stimuli which contradicted their model of the world, as the best response in incongruent trials is to prepare the action opposite to the one the cue points to. Over several trials and in agreement with the instructions they were provided with, the Stroop effect was reduced, suggesting that they adapted their model of the world to the current situation and tried to prepare the opposite action when seeing a cue.

In the studies on cognitive dissonance, we observed that subjects changed their preferences when remembering inconsistencies in their behavior. Preferences can be defined as an indicator of subject's priors. Interestingly, they change when subjects face an inconsistency between their action and this model. It does not rely on the identification of any kind of regularity. Rather, subjects tend to correct any inconsistency in their behavior. This difference between seeking consistency in our environment and in our behavior probably relies on the fact that we do not need to extract patterns in our behavior as we have a privileged access to our model of ourselves and to our preferences. Note that this interpretation contradicts Bem's theory of self-perception (Bem, 1972), as in this theory, the author argues that preferences are inferred based on behavior, and that therefore changes in behavior can directly account for changes in preferences.

3.2.3. Why do we seek consistency?

One last interesting question regarding consistency is why we seek consistency. Indeed, in the introduction of this thesis, I underlined the extent to which interpretations provided by patients or control subjects are consistent and meaningful. Then, I tried to understand the mechanisms associated with the detection and the adaptation to inconsistencies. But I have not yet addressed what can explain this drive.

Looking for regular patterns in the environment can be related to a maximization of predictability and thus a minimization of prediction error. When regularities are identified,

subjects can optimize their prediction about the environment and improve their performance in cognitive tasks, or more generally improve the relevance of their actions. Moreover, as presented in the introduction, subjects tend to identify illusory patterns more readily when they lack control over the situation (Whitson and Galinsky, 2008). It is interesting to notice that these observations resemble the early features of psychosis (Coltheart et al., 2011; Micoulaud-Franchi et al., 2012; Vinckier et al., under review). The identification of regularities in the environment appears as a way to decrease uncertainty that subjects are feeling. Subjective uncertainty has been described as a powerful and uncomfortable experience, that people try to reduce (White, 1959; Hogg, 2000; Corlett et al., 2010). This aversive feeling could be related to an inability to accurately predict the environment and thus to act upon it appropriately. It is interesting to discuss the aversive feeling associated with uncertainty in the environment and the aversive state described by Festinger (1957) as triggered by cognitive dissonance could be related.

The first sentence of the book explaining the cognitive dissonance theory states that "the individual strives toward consistency within himself" (Festinger, 1957). This assumption is not really justified, but rather the cognitive dissonance theory explains this drive as a way to reduce the aversive state arisen by inconsistencies in one's behavior. However, Festinger (1957) proposed no answer to what underlies this aversive state. Harmon-Jones and colleagues addressed this question in their action-based model of dissonance (Harmon-Jones, 1999; Harmon-Jones and Harmon-Jones, 2007). According to this model, perceptions and cognitions could be considered as action tendencies. Inconsistency between two cognitions would evoke an aversive state because it could interfere with effective action. Cognitive dissonance reduction would thus be a way to reduce the conflict between action tendencies and facilitating action execution. In the example of the free-choice paradigm, the positive aspects of the rejected alternative and the negative aspects of the chosen alternative are inconsistent with the decision that has been taken. This inconsistency has to potential of interfering with the translation of the decision into an action. According to the action-based model of dissonance, preference change is a way to commit to the action. Finally, another very speculative account of why we seek internal

consistency as evidenced by cognitive dissonance effects is that it increases the predictability of our own behavior and this could improve communication with others. Indeed, it has been suggested that when interacting with others, we construct models of their minds, similarly to the models we construct about the world (Frith, 2007), and thus we try to optimize these models and reduce prediction errors. Similarly to what we previously discussed about lack of predictability in the environment, uncertainty in social interactions could also trigger aversive feelings that we try to reduce by sampling more information about people we are interacting with, as suggested by some theories of communication (Berger and Calabrese, 1975). Thus, our internal consistency could facilitate our interactions with others.

4. Conclusion

In sum, the studies presented in this thesis addressed the issue of how we handle inconsistencies, both in the environment and in our own behavior, to improve our ability to predict future events and to properly adapt our behavior. The different processes we identified along the four experimental studies are involved in the creation of the complex stories and interpretations, which were the starting point of this thesis. An interesting avenue of future research is to provide a better understanding of the subjective features of these processes, which were only partly addressed in this thesis.

BIBLIOGRAPHY

Abell F, Happé F, Frith U (2000) Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development. Cogn Dev 15:1–16.

Adelson E, Pentland A (1996) The perception of shading and reflectance. In: Perception as Bayesian Inference (Knill DC, Richards W, eds), pp 516. Cambridge University Press.

Alexander MP, Stuss DT, Benson DF (1979) Capgras syndrome: a reduplicative phenomenon. Neurology 29:334–339.

Alho K (1995) Cerebral generators of mismatch negativity (MMN) and its magnetic counterpart (MMNm) elicited by sound changes. Ear Hear 16:38–51.

Alós-Ferrer C, Granić Đ-G, Shi F, Wagner AK (2012) Choices and preferences: Evidence from implicit choices and response times. J Exp Soc Psychol 48:1336–1342.

Alós-Ferrer C, Shi F (2012) Choice-Induced Preference Change: In Defense of the Free-Choice Paradigm. Available SSRN 2062507.

Andersson JLR, Skare S, Ashburner J (2003) How to correct susceptibility distortions in spin-echo echo-planar images: Application to diffusion tensor imaging. Neuroimage 20:870–888.

Ansorge U, Fuchs I, Khalid S, Kunde W (2011) No conflict control in the absence of awareness. Psychol Res 75:351–365.

Ansorge U, Kunde W, Kiefer M (2014) Unconscious vision and executive control: How unconscious processing and conscious action control interact. Conscious Cogn 27C:268–287.

Ansorge U, Neumann O (2005) Intentions determine the effect of invisible metacontrast-masked primes: evidence for top-down contingencies in a peripheral cuing task. J Exp Psychol Hum Percept Perform 31:762–777.

Aronson E (1968) Dissonance theory: Progress and problems. In: Theories of cognitive consistency: A sourcebook (Abelson RP, Aronson E, Mcguire WJ, Newcomb TM, Rosenberg MJ, Tannenbaum PH, eds), pp 5–27. Rand McNally.

Aronson E, Mills J (1959) The effect of severity of initiation on liking for a group. J Abnorm Soc Psychol 59:177–181.

Aru J, Bachmann T, Singer W, Melloni L (2012) Distilling the neural correlates of consciousness. Neurosci Biobehav Rev 36:737–746.

Atienza M, Cantero JL, Gómez CM (1997) The mismatch negativity component reveals the sensory memory during REM sleep in humans. Neurosci Lett 237:21–24.

Axmacher N, Cohen MX, Fell J, Haupt S, Dümpelmann M, Elger CE, Schlaepfer TE, Lenartz D, Sturm V, Ranganath C (2010) Intracranial EEG correlates of expectancy and memory formation in the human hippocampus and nucleus accumbens. Neuron 65:541–549.

Baayen RH, Davidson DJ, Bates DM (2008) Mixed-effects modeling with crossed random effects for subjects and items. J Mem Lang 59:390–412.

Banich MT, Milham MP, Atchley R, Cohen NJ, Webb A, Wszalek T, Kramer AF, Liang Z, Wright A, Shenker J, Magin R (1991) fMRI Studies of Stroop Tasks Reveal Unique Roles of Anterior and Posterior Brain Systems in Attentional Selection. :988–1000.

Barch D, Braver T, Akbudak E, Conturo T, Ollinger J, Snyder A (2001) Anterior cingulate cortex and response conflict: effects of response modality and processing domain. Cereb Cortex 11:837–848.

Bartra O, McGuire JT, Kable JW (2013) The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. Neuroimage 76:412–427.

Baudena P, Halgren E, Heit G, Clarke JM (1995) Intracerebral potentials to rare target and distractor auditory and visual stimuli. III. Frontal cortex. Electroencephalogr Clin Neurophysiol 94:251–264.

Bekinschtein TA, Dehaene S, Rohaut B, Tadel F, Cohen L, Naccache L (2009) Neural signature of the conscious processing of auditory regularities. Proc Natl Acad Sci U S A 106:1672–1677.

Bem D (1972) Self-perception theory. Adv Exp Soc Psychol 6:1–62.

Bench C, Frith C, Grasby P, Friston K, Paulesu E, Frackowiak R, Dolan R (1993) Investigations of the functional anatomy of attention using the Stroop test. Neuropsychologia 31:907–922.

Berger C, Calabrese R (1975) Some explorations in initial interaction and beyond: Toward a developmental theory of interpersonal communication. Hum Commun Res 1:99–112.

Bodner GE, Masson MEJ (2001) Prime Validity Affects Masked Repetition Priming: Evidence for an Episodic Resource Account of Priming. J Mem Lang 45:616–647.

Bodner GE, Masson MEJ (2004) Beyond binary judgments: prime validity modulates masked repetition priming in the naming task. Mem Cognit 32:1–11.

Bodner GE, Mulji R (2010) Prime proportion affects masked priming of fixed and free-choice responses. Exp Psychol 57:360–366.

Bonini F, Burle B, Liégeois-Chauvel C, Régis J, Chauvel P, Vidal F (2014) Action monitoring and medial frontal cortex: leading role of supplementary motor area. Science 343:888–891.

Botvinick MM, Braver TS, Barch DM, Carter CS, Cohen JD (2001) Conflict monitoring and cognitive control. Psychol Rev 108:624–652.

Botvinick MM, Cohen JD, Carter CS (2004) Conflict monitoring and anterior cingulate cortex: an update. Trends Cogn Sci 8:539–546.

Bowler DM, Thommen E (2000) Attribution of mechanical and social causality to animated displays by children with. Autism 4:147–171.

Brehm JW (1956) Postdecision changes in the desirability of alternatives. J Abnorm Soc Psychol 52:384–389.

Burgess PW, Shallice T (1996) Confabulation and the Control of Recollection. Memory 4:359–411.

Capgras J, Reboul-Lachaux J (1923) L'illusion des" sosies" dans un délire systématisé chronique. Bull Soc Clin Med Ment 11:6–16.

Carter C, Macdonald AM, Botvinick MM, Ross LL, Stenger VA, Noll D, Cohen JD (2000) Parsing executive processes: strategic vs. evaluative functions of the anterior cingulate cortex. Proc Natl Acad Sci 97:1944–1948.

Castelli F, Frith C, Happé F, Frith U (2002) Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. Brain 125:1839–1849.

Chen MK, Risen JL (2010) How choice affects and reflects preferences: revisiting the free-choice paradigm. J Pers Soc Psychol 99:573–594.

Chennu S, Noreika V, Gueorguiev D, Blenkmann A, Kochen S, Ibáñez A, Owen AM, Bekinschtein T a (2013) Expectation and attention in hierarchical auditory prediction. J Neurosci 33:11194–11205.

Chun MM, Jiang Y (1998) Contextual cueing: implicit learning and memory of visual context guides spatial attention. Cogn Psychol 36:28–71.

Chun MM, Turk-Browne NB (2007) Interactions between attention and memory. Curr Opin Neurobiol 17:177–184.

Clark A (2013) Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behav Brain Sci 36:181–204.

Clayson PE, Larson MJ (2011) Conflict adaptation and sequential trial effects: support for the conflict monitoring theory. Neuropsychologia 49:1953–1961.

Coltheart M, Langdon R, McKay R (2011) Delusional belief. Annu Rev Psychol 62:271–298.

Coltheart M, Menzies P, Sutton J (2010) Abductive inference and delusional belief. Cogn Neuropsychiatry 15:261–287.

Cooper J (2007) Cognitive Dissonance: 50 Years of a Classic Theory. SAGE Publi.

Cooper J, Fazio RH (1984) A new look at dissonance theory. Adv Exp Soc Psychol 17:229–266.

Coppin G, Delplanque S, Cayeux I, Porcherot C, Sander D (2010) I'm no longer torn after choice: how explicit choices implicitly shape preferences of odors. Psychol Sci 21:489–493.

Corlett PR, Taylor JR, Wang X-J, Fletcher PC, Krystal JH (2010) Toward a neurobiology of delusions. Prog Neurobiol 92:345–369.

Damasio AR, Tranel D, Damasio H (1990) Individuals with sociopathic behavior caused by frontal damage fail to respond autonomically to social stimuli. Behav Brain Res 41:81–94.

Danielmeier C, Wessel JR, Steinhauser M, Ullsperger M (2009) Modulation of the error-related negativity by response conflict. Psychophysiology 46:1288–1298.

Darwin C, Turvey M, Crowder R (1972) An auditory analogue of the Sperling partial report procedure: Evidence for brief auditory storage. Cogn Psychol 3:255–267.

Daza MT, Ortells JJ, Fox E (2002) Perception without awareness: further evidence from a Stroop priming task. Percept Psychophys 64:1316–1324.

De Moortel I, Munday SA, Hood AW (2004) Wavelet Analysis: the effect of varying basic wavelet parameters. Sol Phys 222:203–228.

Dehaene S, Changeux J-P (2011) Experimental and theoretical approaches to conscious processing. Neuron 70:200–227.

Dehaene S, Changeux J-P, Naccache L, Sackur J, Sergent C (2006) Conscious, preconscious, and subliminal processing: a testable taxonomy. Trends Cogn Sci 10:204–211.

Dehaene S, Kerszberg M, Changeux J-P (1998) A neuronal model of a global workspace in effortful cognitive tasks. Proc Natl Acad Sci U S A 95:14529–14534.

Dehaene S, Naccache L (2001) Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. Cognition 79:1–37.

Dennett DC (1992) Consciousness Explained. Back Bay Books.

Desender K, Van den Bussche E (2012) Is Consciousness Necessary for Conflict Adaptation? A State of the Art. Front Hum Neurosci 6:1–13.

Desender K, Van Lierde E, Van den Bussche E (2013) Comparing conscious and unconscious conflict adaptation. PLoS One 8:e55976.

Desender K, Van Opstal F, Van den Bussche E (2014) Feeling the conflict: the crucial role of conflict experience in adaptation. Psychol Sci 25:675–683.

Donchin E, Coles MGH (1988) Is the P300 component a manifestation of context updating? Behav Brain Sci 11:357–374.

Donner TH, Siegel M (2011) A framework for local cortical oscillation patterns. Trends Cogn Sci 15:191–199.

Dupoux E, de Gardelle V, Kouider S (2008) Subliminal speech perception and auditory streaming. Cognition 109:267–273.

Edwards E, Soltani M, Deouell LY, Berger MS, Knight RT (2005) High gamma activity in response to deviant auditory stimuli recorded directly from human cortex. J Neurophysiol 94:4269–4280.

Egan LC, Bloom P, Santos LR (2010) Choice-induced preferences in the absence of choice: Evidence from a blind two choice paradigm with young children and capuchin monkeys. J Exp Soc Psychol 46:204–207.

Egan LC, Santos LR, Bloom P (2007) The origins of cognitive dissonance: evidence from children and monkeys. Psychol Sci 18:978–983.

Eichenbaum H, Yonelinas a P, Ranganath C (2007) The medial temporal lobe and recognition memory. Annu Rev Neurosci 30:123–152.

Ellis HD, Young AW, Quayle AH, De Pauw KW (1997) Reduced autonomic responses to faces in Capgras delusion. Proc Biol Sci 264:1085–1092.

Eriksen B, Eriksen C (1974) Effects of noise letters upon the identification of a target letter in a nonsearch task. Percept Psychophys 16:143–149.

Escera C, Alho K, Winkler I, Näätänen R (1998) Neural mechanisms of involuntary attention to acoustic novelty and change. J Cogn Neurosci 10:590–604.

Escera C, Yago E, Alho K (2001) Electrical responses reveal the temporal dynamics of brain events during involuntary attention switching. Eur J Neurosci 14:877–883.

Escera C, Yago E, Corral M-J, Corbera S, Nunez MI (2003) Attention capture by auditory significant stimuli: semantic analysis follows attention switching. Eur J Neurosci 18:2408–2412.

Fang Q (2010) Mesh-based Monte Carlo method using fast ray-tracing in Plücker coordinates. Biomed Opt Express 1:165–175.

Fang Q, Boas DA (2009) Tetrahedral mesh generation from volumetric binary and grayscale images. IEEE Int Symp Biomed Imaging 2009:1142–1145.

Faugeras F, Rohaut B, Weiss N, Bekinschtein T, Galanaud D, Puybasset L, Bolgert F, Sergent C, Cohen L, Dehaene S, Naccache L (2012) Event related potentials elicited by violations of auditory regularities in patients with impaired consciousness. Neuropsychologia 50:403–418.

Faugeras F, Rohaut B, Weiss N, Bekinschtein TA, Galanaud D, Puybasset L, Bolgert F, Sergent C, Cohen L, Dehaene S, Naccache L (2011) Probing consciousness with event-related potentials in the vegetative state. Neurology 77:264–268.

Feinberg TE, Keenan JP (2005) Where in the brain is the self? Conscious Cogn 14:661–678.

Festinger L (1957) A theory of cognitive dissonance. Stanford Univ Pr.

Festinger L, Carlsmith JM (1959) Cognitive consequences of forced compliance. J Abnorm Soc Psychol 58:203–210.

Festinger L, Riecken HW, Schachter S (1956) When prophecy fails. University of Minnesota Press.

Fischer C, Morlet D, Bouchet P, Luaute J, Jourdan C, Salord F (1999) Mismatch negativity and late auditory evoked potentials in comatose patients. Clin Neurophysiol 110:1601–1610.

Fletcher PC, Frith CD (2009) Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. Nat Rev Neurosci 10:48–58.

Folstein JR, Van Petten C (2008) Influence of cognitive control and mismatch on the N2 component of the ERP: a review. Psychophysiology 45:152–170.

Forster SE, Carter CS, Cohen JD, Cho RY (2010) Parametric Manipulation of the Conflict Signal and Control-state Adaptation. J Cogn Neurosci 23:923–935.

Friedman D, Cycowicz YM, Gaeta H (2001) The novelty P3: an event-related brain potential (ERP) sign of the brain's evaluation of novelty. Neurosci Biobehav Rev 25:355–373.

Fries A, Frey D (1980) Misattribution of arousal and the effects of self-threatening information. J Exp Soc Psychol 16:405–416.

Fries P (2005) A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. Trends Cogn Sci 9:474–480.

Friston K (2005) A theory of cortical responses. Philos Trans R Soc Lond B Biol Sci 360:815–836.

Friston K (2010) The free-energy principle: a unified brain theory? Nat Rev Neurosci 11:127–138.

Frith C (2007) Making up the Mind: How the Brain Creates Our Mental World. Wiley-Blackwell.

Frühholz S, Godde B, Finke M, Herrmann M (2011) Spatio-temporal brain dynamics in a combined stimulus-stimulus and stimulus-response conflict task. Neuroimage 54:622–634.

Gaillard R, Dehaene S, Adam C, Clémenceau S, Hasboun D, Baulac M, Cohen L, Naccache L (2009) Converging Intracranial Markers of Conscious Access. PLoS Biol 7:e1000061.

Garrido MI, Friston KJ, Kiebel SJ, Stephan KE, Baldeweg T, Kilner JM (2008) The functional anatomy of the MMN: a DCM study of the roving paradigm. Neuroimage 42:936–944.

Garrido MI, Kilner JM, Stephan KE, Friston KJ (2009) The mismatch negativity: a review of underlying mechanisms. Clin Neurophysiol 120:453–463.

Gazzaniga MS (1967) The Split Brain in Man. Sci Am 217:24–29.

Gazzaniga MS (1989) Organization of the human brain. Science (80-) 245:947–952.

Gazzaniga MS (1998) The split brain revisited. Sci Am.

Gazzaniga MS (2000) Cerebral specialization and interhemispheric communication Does the corpus callosum enable the human condition? Brain 123:1293–1326.

Gazzaniga MS, Bogen J, Sperry R (1965) Observations on visual perception after disconnexion of the cerebral hemispheres in man. Brain 88.

Gazzaniga MS, LeDoux J, Wilson D (1977) Language, praxis, and the right hemisphere Clues to some mechanisms of consciousness. Neurology:1144–1147.

Gergely G, Nadasdy Z, Csibra G, Bfr S (1995) Taking the intentional stance at 12 months of age. Cognition 56:165–193.

Ghinescu R, Schachtman TR, Stadler M a, Fabiani M, Gratton G (2010) Strategic behavior without awareness? Effects of implicit learning in the Eriksen flanker paradigm. Mem Cognit 38:197–205.

Giard M-H, Perrin F, Pernier J, Bouchet P (1990) Brain Generators Implicated in the Processing of Auditory Stimulus Deviance: A Topographic Event-Related Potential Study. Psychophysiology 27:627–640.

Gilboa A, Moscovitch M (2002) The Cognitive Neuroscience of Confabulation: A Review and a Model. In: The Handbook of Memory Disorders, 2nd ed. (Baddeley AD, Kopelman MD, Wilson BA, eds), pp 315–342. Chichester: John Wiley & Sons, Ltd.

Gratton G, Coles MG, Donchin E (1992) Optimizing the use of information: strategic control of activation of responses. J Exp Psychol Gen 121:480–506.

Greenwald A, Draine S, Abrams R (1996) Three cognitive markers of unconscious semantic activation. Science (80-) 273:1699–1702.

Greenwald AG, Klinger MR, Schuh ES (1995) Activation by marginally perceptible ("subliminal") stimuli: Dissociation of unconscious from conscious cognition. J Exp Psychol Gen 124:22–42.

Grimm S, Escera C, Slabu L, Costa-Faidella J (2011) Electrophysiological evidence for the hierarchical organization of auditory change detection in the human brain. Psychophysiology 48:377–384.

Halgren E, Baudena P, Clarke JM, Heit G, Liégeois C, Chauvel P, Musolino A (1995a) Intracerebral potentials to rare target and distractor auditory and visual stimuli. I. Superior temporal plane and parietal lobe. Electroencephalogr Clin Neurophysiol 94:191–220.

Halgren E, Baudena P, Clarke JM, Heit G, Marinkovic K, Devaux B, Vignal J-P, Biraben A (1995b) Intracerebral potentials to rare target and distractor auditory and visual stimuli. II. Medial, lateral and posterior temporal lobe. Electroencephalogr Clin Neurophysiol 94:229–250.

Halgren E, Marinkovic K, Chauvel P (1998) Generators of the late cognitive potentials in auditory and visual oddball tasks. Electroencephalogr Clin Neurophysiol 106:156–164.

Hanslmayr S, Pastötter B, Bäuml K-H, Gruber S, Wimber M, Klimesch W (2008) The electrophysiological dynamics of interference during the Stroop task. J Cogn Neurosci 20:215–225.

Harmon-Jones E (1999) Towards an understanding of the motivation underlying dissonance effects: is the production of aversive consequences necessary? In: Cognitive Dissonance: Perspectives on a pivotal theory in social psychology (Harmon-Jones E, Mills J, eds), pp 71–99. Washington, DC: American Psychological Association.

Harmon-Jones E, Harmon-Jones C (2007) Cognitive Dissonance Theory After 50 Years of Development. Zeitschrift für Sozialpsychologie 38:7–16.

Heider F, Simmel M (1944) An experimental study of apparent behavior. Am J Psychol 57:243–259.

Heine S, Dehman D (1997) Culture, dissonance, and self-affirmation. Personal Soc Psychol Bull 23:389–400.

Heine SJ, Proulx T, Vohs KD (2006) The meaning maintenance model: on the coherence of social motivations. Pers Soc Psychol Rev 10:88–110.

Helmholtz H von (1910) Helmholtz's Treatise on Physiological Optics (vol. 3), 1910 trans. The Optical Society of America and University of Pennsylvania.

Hinson JM, Staddon JER (1983) Matching, maximizing and hill-climbing. J Exp Anal Behav 40:321–331.

Hirsh JB, Mar RA, Peterson JB (2013) Personal narratives as the highest level of cognitive integration. Behav Brain Sci 36:216–217.

Hirstein W, Ramachandran VS (1997) Capgras syndrome: a novel probe for understanding the neural representation of the identity and familiarity of persons. Proc R Soc London 264:437–444.

Hogg MA (2000) Subjective Uncertainty Reduction through Self-categorization: A Motivational Theory of Social Identity Processes. Eur J Soc Psychol 11:223–255.

Holden S (2013) Do Choices Affect Preferences ? Some Doubts and New Evidence. J Appl Soc Psychol 43:83–94.

Izuma K, Matsumoto M, Murayama K, Samejima K, Sadato N, Matsumoto K (2010) Neural correlates of cognitive dissonance and choice-induced preference change. Proc Natl Acad Sci U S A 107:22014–22019.

Izuma K, Murayama K (2013) Choice-induced preference change in the free-choice paradigm: a critical methodological review. Front Psychol 4.

Jääskeläinen I, Ahveninen J, Bonmassar G, Dale AM, Ilmoniemi RJ, Levänen S, Lin F-H, May P, Melcher J, Stufflebeam S, Tiitinen H, Belliveau JW (2004) Human posterior auditory cortex gates novel sounds to consciousness. Proc Natl Acad Sci 101:6809–6814.

Jack a I, Shallice T (2001) Introspective physicalism as an approach to the science of consciousness. Cognition 79:161–196.

Jarcho JM, Berkman ET, Lieberman MD (2011) The neural basis of rationalization: Cognitive dissonance reduction during decision-making. Soc Cogn Affect Neurosci 6:460–467.

Jáskowski P, Skalska B, Verleger R (2003) How the self controls its "automatic pilot" when processing subliminal information. J Cogn Neurosci 15:911–920.

Jiang J, van Gaal S, Bailey K, Chen A, Zhang Q (2013) Electrophysiological correlates of block-wise strategic adaptations to consciously and unconsciously triggered conflict. Neuropsychologia 51:2791–2798.

Johansson P, Hall L, Sikström S, Olsson A (2005) Failure to detect mismatches between intention and outcome in a simple decision task. Science 310:116–119.

Johansson P, Hall L, Tärning B, Sikström S, Chater N (2013) Choice Blindness and Preference Change: You Will Like This Paper Better If You (Believe You) Chose to Read It! J Behav Decis Mak:n/a – n/a.

Kable JW, Glimcher PW (2009) The Neurobiology of Decision: Consensus and Controversy. Neuron 63:733–745.

Kekoni J, Hämäläinen H, Saarinen M, Gröhn J, Reinikainen K, Lehtokoski A, Näätänen R (1997) Rate effect and mismatch responses in the somatosensory system: ERP-recordings in humans. Biol Psychol 46:125–142.

Kentridge R (1999) Attention without awareness in blindsight. Proc R Soc London Biol 266:1805–1811.

Kerns JG, Cohen JD, MacDonald AW, Cho RY, Stenger VA, Carter CS (2004) Anterior cingulate conflict monitoring and adjustments in control. Science (80-) 303:1023–1026.

Kiefer M (2012) Executive control over unconscious cognition: attentional sensitization of unconscious information processing. Front Hum Neurosci 6:61.

Kiefer M, Brendel D (2006) Attentional modulation of unconscious "automatic" processes: evidence from event-related potentials in a masked priming paradigm. J Cogn Neurosci 18:184–198.

Kim H, Adolphs R, O'Doherty JP, Shimojo S (2007) Temporal isolation of neural processes underlying face preference decisions. Proc Natl Acad Sci 104:18253–18258.

King J (2014) Characterizing the electro-magnetic signatures of conscious processing in healthy and impaired human brains. :258.

King JR, Faugeras F, Gramfort A, Schurger A, El Karoui I, Sitt JD, Rohaut B, Wacongne C, Labyt E, Bekinschtein T, Cohen L, Naccache L, Dehaene S (2013) Single-trial decoding of auditory novelty responses facilitates the detection of residual consciousness. Neuroimage 83:726–738.

Kinoshita S, Forster KI, Mozer MC (2008) Unconscious cognition isn't that smart: modulation of masked repetition priming effect in the word naming task. Cognition 107:623–649.

Kitayama S, Chua HF, Tompson S, Han S (2013) Neural mechanisms of dissonance: An fMRI investigation of choice justification. Neuroimage 69:206–212.

Klapp ST (2007) Nonconscious control mimics a purposeful strategy: strength of Stroop-like interference is automatically modulated by proportion of compatible trials. J Exp Psychol Hum Percept Perform 33:1366–1376.

Knill DC, Pouget A (2004) The Bayesian brain: the role of uncertainty in neural coding and computation. Trends Neurosci 27:712–719.

Koelsch S, Heinke W, Sammler D, Olthoff D (2006) Auditory processing during deep propofol sedation and recovery from unconsciousness. Clin Neurophysiol 117:1746–1759.

Kok A (2001) On the utility of P3 amplitude as a measure of processing capacity. Psychophysiology 38:557–577.

Kopell NJ, Ermentrout GB, Whittington M, Traub RD (2000) Gamma rhythms and beta rhythms have different synchronization properties. Proc Natl Acad Sci U S A 97:1867–1872.

Kopp B, Rist F, Mattler U (1996) N200 in the flanker task as a neurobehavioral tool for investigating executive control. Psychophysiology 33:282–294.

Korsakoff S (1889) Etude médico-psychologique sur une forme des maladies de la mémoire. Rev Philos:1–19.

Korsakoff S (1955) Psychic disorder in conjunction with peripheral neuritis (translated by M. Victor and P.I. Yakovlev). Neurology 5:394–406.

Kropotov JD, Alho K, Näätänen R, Ponomarev V a, Kropotova O V, Anichkov a D, Nechaev VB (2000) Human auditory-cortex mechanisms of preattentive sound discrimination. Neurosci Lett 280:87–90.

Kunde W (2003) Sequential modulations of stimulus-response correspondence effects depend on awareness of response conflict. Psychon Bull Rev 10:198–205.

Kunde W, Reuss H, Kiesel A (2012) Consciousness and cognitive control. Adv Cogn Psychol 8:9–18.

Lachaux J-P, Rodriguez E, Martinerie J, Varela FJ (1999) Measuring Phase Synchrony in Brain Signals. Hum Brain Mapp 8:194–208.

Langdon R, Coltheart M (2000) The Cognitive Neuropsychology of Delusions. Mind Lang 15:184–218.

Lau HC, Passingham RE (2007) Unconscious activation of the cognitive control system in the human prefrontal cortex. J Neurosci 27:5805–5811.

Lebreton M, Jorge S, Michel V, Thirion B, Pessiglione M (2009) An automatic valuation system in the human brain: evidence from functional neuroimaging. Neuron 64:431–439.

Levy I, Lazzaro SC, Rutledge RB, Glimcher PW (2011) Choice from non-choice: predicting consumer preferences from blood oxygenation level-dependent signals obtained during passive viewing. J Neurosci 31:118–125.

Liasis a, Towell A, Alho K, Boyd S (2001) Intracranial identification of an electric frontal-cortex response to auditory stimulus change: a case study. Brain Res Cogn Brain Res 11:227–233.

Lieberman MD, Ochsner KN, Gilbert DT, Schacter DL (2001) Do Amnesics Exhibit Cognitive Dissonance Reduction? The Role of Explicit Memory and Attention in Attitude Change. Psychol Sci 12:135–140.

Liu Z, Knill DC, Kersten D (1995) Object classification for human and ideal observers. Vision Res 35:549–568.

Logan G (1982) On the ability to inhibit complex movements: A stop-signal study of typewriting. J Exp Psychol Hum Percept Perform 8:778–792.

Logan G, Zbrodoff N (1979) When it helps to be misled: Facilitative effects of increasing the frequency of conflicting stimuli in a Stroop-like task. Mem Cognit 7:166–174.

Maier NRF (1931) Reasoning in humans. II. The solution of a problem and its appearance in consciousness. J Comp Psychol 12:181–194.

Maldjian JA, Laurienti PJ, Burdette JH (2004) Precentral gyrus discrepancy in electronic versions of the Talairach atlas. Neuroimage 21:450–455.

Maldjian JA, Laurienti PJ, Kraft RA, Burdette JH (2003) An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. Neuroimage 19:1233–1239.

Mamassian P, Landy MS (2001) Interaction of visual prior constraints. Vision Res 41:2653–2668.

Maris E, Oostenveld R (2007) Nonparametric statistical testing of EEG- and MEG-data. J Neurosci Methods 164:177–190.

Marzouki Y, Grainger J, Theeuwes J (2007) Exogenous spatial cueing modulates subliminal masked priming. Acta Psychol (Amst) 126:34–45.

May PJC, Tiitinen H (2010) Mismatch negativity (MMN), the deviance-elicited auditory deflection, explained. Psychophysiology 47:66–122.

Mayr U, Awh E, Laurey P (2003) Conflict adaptation effects in the absence of executive control. Nat Neurosci 6:450–452.

Mengarelli F, Spoglianti S, Avenanti A, di Pellegrino G (2013) Cathodal tDCS Over the Left Prefrontal Cortex Diminishes Choice-Induced Preference Change. Cereb Cortex.

Merikle PM, Cheesman J (1987) Current Status of Research on Subliminal Perception. Adv Consum Res 14:298–302.

Merikle PM, Joordens S (1997) Parallels between Perception without Attention and Perception without Awareness. Conscious Cogn 6:219–236.

Merikle PM, Joordens S, Stolz JA (1995) Measuring the relative magnitude of unconscious influences. Conscious Cogn 4:422–439.

Metcalf K, Langdon R, Coltheart M (2007) Models of confabulation: a critical review and a new framework. Cogn Neuropsychol 24:23–47.

Micoulaud-Franchi J-A, Aramaki M, Merer A, Cermolacce M, Ystad S, Kronland-Martinet R, Naudin J, Vion-Dury J (2012) Toward an exploration of feeling of strangeness in schizophrenia: perspectives on acousmatic and everyday listening. J Abnorm Psychol 121:628–640.

Molholm S, Martinez A, Ritter W, Javitt DC, Foxe JJ (2005) The neural circuitry of pre-attentive auditory change-detection: an fMRI study of pitch and duration mismatch negativity generators. Cereb cortex 15:545–551.

Monsell S (2003) Task switching. Trends Cogn Sci 7:134–140.

Näätänen R, Gaillard AWK (1983) The orienting reflex and the N2 deflection of the event-related potentials. In: Tutorials in Event Related Potential Research: Endogenous Components, pp 119–141 Advances in Psychology. Elsevier.

Näätänen R, Gaillard AWK, Mäntysalo S (1978) Early selective-attention effect on evoked potential reinterpreted. Acta Psychol (Amst) 42:313–329.

Näätänen R, Paavilainen P, Rinne T, Alho K (2007) The mismatch negativity (MMN) in basic research of central auditory processing: a review. Clin Neurophysiol 118:2544–2590.

Näätänen R, Tervaniemi M, Sussman E, Paavilainen P, Winkler I (2001) "Primitive intelligence" in the auditory cortex. Trends Neurosci 24:283–288.

Naccache L (2006) Le Nouvel inconscient : Freud, le Christophe Colomb des neurosciences. Paris: Odile Jacob.

Naccache L, Blandin E, Dehaene S (2002) Unconscious masked priming depends on temporal attention. Psychol Sci 13:416–424.

Naccache L, Puybasset L, Gaillard R, Serve E, Willer J-C (2005) Auditory mismatch negativity is a good predictor of awakening in comatose patients: a fast and reliable procedure. Clin Neurophysiol 116:988–989.

Naccache L, Sportiche S, Strauss M, El Karoui I, Sitt J, Cohen L (2014) Imaging "top-down" mobilization of visual information: a case study in a posterior split-brain patient. Neuropsychologia 53:94–103.

Nakamura K, Kawabata H (2013) I choose, therefore I like: preference for faces induced by arbitrary choice. PLoS One 8:e72071.

Nichols TE, Holmes AP (2002) Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum Brain Mapp 15:1–25.

Nieuwenhuis S, Yeung N, van den Wildenberg W, Ridderinkhof KR (2003) Electrophysiological correlates of anterior cingulate function in a go/no-go task: Effects of response conflict and trial type frequency. Cogn Affect Behav Neurosci 3:17–26.

Nir Y, Fisch L, Mukamel R, Gelbard-Sagiv H, Arieli A, Fried I, Malach R (2007) Coupling between Neuronal Firing Rate , Gamma LFP , and BOLD fMRI Is Related to Interneuronal Correlations. Curr Biol 17:1275–1285.

Nisbett RE, Wilson TD (1977) Telling more than we can know: Verbal reports on mental processes. Psychol Rev 84:231.

Nissen M, Bullemer P (1987) Attentional requirements of learning: Evidence from performance measures. Cogn Psychol 19:1–32.

Oostenveld R, Fries P, Maris E, Schoffelen J-M (2011) FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. Comput Intell Neurosci 2011:1–9.

Opitz B, Rinne T, Mecklinger A, von Cramon DY, Schröger E (2002) Differential contribution of frontal and temporal cortices to auditory change detection: fMRI and ERP results. Neuroimage 15:167–174.

Ortells JJ, Fox E, Noguera C, Abad MJ (2003) Repetition priming effects from attended vs . ignored single words in a semantic categorization task. Acta Psychol (Amst) 114:185–210.

Pardo J V., Pardo PJ, Janer KW, Raichle ME (1990) The anterior cingulate cortex mediates processing selection in the Stroop attentional conflict paradigm. Proc Natl Acad Sci 87:256–259.

Pause B, Krauel K (2000) Chemosensory event-related potentials (CSERP) as a key to the psychology of odors. Int J Psychophysiol 36:105–122.

Pazo-Alvarez P, Cadaveira F, Amenedo E (2003) MMN in the visual modality: a review. Biol Psychol 63:199–236.

Pegado F, Bekinschtein TA, Chausson N, Dehaene S, Cohen L, Naccache L (2010) Probing the lifetimes of auditory novelty detection processes. Neuropsychologia 48:3145–3154.

Pesaran B, Pezaris JS, Sahani M, Mitra PP, Andersen R a (2002) Temporal structure in neuronal activity during working memory in macaque parietal cortex. Nat Neurosci 5:805–811.

Pfurtscheller G, Lopes da Silva FH (1999) Event-related EEG/MEG synchronization and desynchronization: basic principles. Clin Neurophysiol 110:1842–1857.

Picton TW (1992) The P300 wave of the human event-related potential. J Clin Neurophysiol 9:456–479.

Polich J (2007) Updating P300: an integrative theory of P3a and P3b. Clin Neurophysiol 118:2128–2148.

Preston AR, Eichenbaum H (2013) Interplay of hippocampus and prefrontal cortex in memory. Curr Biol 23:R764–R773.

Proulx T, Inzlicht M (2012) The Five “A”s of Meaning Maintenance: Finding Meaning in the Theories of Sense-Making. Psychol Inq 23:317–335.

Proulx T, Inzlicht M, Harmon-Jones E (2012) Understanding all inconsistency compensation as a palliative response to violated expectations. Trends Cogn Sci 16:285–291.

Qin J, Kimel S, Kitayama S, Wang X, Yang X, Han S (2011) How choice modifies preference : Neural correlates of choice justification. Neuroimage 55:240–246.

Ramachandran V (1995) Anosognosia in parietal lobe syndrome. Conscious Cogn 4:22–51.

Ranganath C, Rainer G (2003) Neural mechanisms for detecting and remembering novel events. Nat Rev Neurosci 4:193–202.

Rao R, Ballard D (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci 2:79–87.

Ray S, Maunsell JHR (2010) Differences in gamma frequencies across visual cortex restrict their possible use in computation. Neuron 67:885–896.

Ridderinkhof KR, Ullsperger M, Crone E a, Nieuwenhuis S (2004a) The role of the medial frontal cortex in cognitive control. Science 306:443–447.

Ridderinkhof KR, van den Wildenberg WPM, Segalowitz SJ, Carter CS (2004b) Neurocognitive mechanisms of cognitive control: the role of prefrontal cortex in action selection, response inhibition, performance monitoring, and reward-based learning. Brain Cogn 56:129–140.

Rinne T, Alho K, Ilmoniemi RJ, Virtanen J, Näätänen R (2000) Separate time behaviors of the temporal and frontal mismatch negativity sources. Neuroimage 12:14–19.

Rinne T, Degerman A, Alho K (2005) Superior temporal and inferior frontal cortices are activated by infrequent sound duration decrements: an fMRI study. Neuroimage 26:66–72.

Risen JL, Chen MK (2010) How to Study Choice-Induced Attitude Change : Strategies for Fixing the Free-Choice Paradigm. Soc Personal Psychol Compass 12:1151–1164.

Rosburg T, Trautner P, Dietl T, Korzyukov OA, Boutros NN, Schaller C, Elger CE, Kurthen M (2005) Subdural recordings of the mismatch negativity (MMN) in patients with focal epilepsy. Brain 128:819–828.

Saffran J, Aslin R, Newport E (1996) Statistical learning by 8-month-old infants. Science (80-) 274:1926–1928.

Salti M, El Karoui I, Maillet M, Naccache L (2014) Cognitive Dissonance Resolution Is Related to Episodic Memory. PLoS One 9:1–8.

Sato Y, Yabe H, Hiruma T, Sutoh T, Shinozaki N, Nashida T, Kaneko S (2000) The effect of deviant stimulus probability on the human mismatch process. Neuroreport 11:3703–3708.

Schacter DL, Norman KA, Koutstaal W (1998) The cognitive neuroscience of constructive memory. Annu Rev Psychol 49:289–318.

Schnider A (2003) Spontaneous confabulation and the adaptation of thought to ongoing reality. Nat Rev Neurosci 4:662–671.

Schnider A (2013) Orbitofrontal reality filtering. Front Behav Neurosci 7:67.

Sergent C, Baillet S, Dehaene S (2005) Timing of the brain events underlying access to consciousness during the attentional blink. Nat Neurosci 8:1391–1400.

Sergent C, Naccache L (2012) Imaging neural signatures of consciousness: “what”, “when”, “where” and “how” does it work? Arch Ital Biol 150:91–106.

Sergent C, Wyart V, Babo-Rebelo M, Cohen L, Naccache L, Tallon-Baudry C (2013) Cueing attention after the stimulus is gone can retrospectively trigger conscious perception. Curr Biol 23:150–155.

Sharot T, Fleming SM, Yu X, Koster R, Dolan RJ (2012) Is Choice-Induced Preference Change Long Lasting? Psychol Sci 23:1123–1129.

Sharot T, Martino B De, Dolan RJ (2009) How Choice Reveals and Shapes Expected Hedonic Outcome. J Neurosci 29:3760–3765.

Sharot T, Velasquez CM, Dolan RJ (2010) Do decisions shape preference?: Evidence from blind choice. Psychol Sci 21:1231–1235.

Sherman DK, Cohen GVL (2006) The psychology of self-defense: self-affirmation theory. Adv Exp Soc Psychol 38:183–242.

Simon L, Greenberg J, Brehm J (1995) Trivialization: the forgotten mode of dissonance reduction. J Pers Soc Psychol 68:247–260.

Sitt JD, King J-R, El Karoui I, Rohaut B, Faugeras F, Gramfort A, Cohen L, Sigman M, Dehaene S, Naccache L (2014) Large scale screening of neural signatures of consciousness in patients in a vegetative or minimally conscious state. Brain 137:2258–2270.

Smith ME, Halgren E, Sokolik M, Baudena P, Musolino A, Liegeois-Chauvel C, Chauvel P (1990) The intracranial topography of the P3 event-related potential elicited during auditory oddball. Electroencephalogr Clin Neurophysiol 76:235–248.

Squire LR (1982) The neuropsychology of human memory. Annu Rev Neurosci 5:241–273.

Squires NK, Squires KC, Hillyard SA (1975) Two varieties of long-latency positive waves evoked by unpredictable auditory stimuli in man. Electroencephalogr Clin Neurophysiol 38:387–401.

Steele CM (1988) The Psychology of Self-Affirmation: Sustaining the Integrity of the Self. Adv Exp Soc Psychol 21:261–302.

Steele CM, Spencer SJ, Lynch M (1993) Self-image resilience and dissonance: The role of affirmational resources. J Pers Soc Psychol 64:885–896.

Sternberg S (2001) Separate modifiability, mental modules, and the use of pure and composite measures to reveal them. Acta Psychol (Amst) 106:147–246.

Stone J, Cooper J (2001) A self-standards model of cognitive dissonance. J Exp Soc Psychol 37:228–243.

Stroop J (1935) Studies of interference in serial verbal reactions. J Exp Psychol:1–8.

Summerfield C, Koechlin E (2008) A neural representation of prior information during perceptual inference. Neuron 59:336–347.

Sussman E, Winkler I (2001) Dynamic sensory updating in the auditory system. Cogn Brain Res 12:431–439.

Sutton S, Braren M, Zubin J, John ER (1965) Evoked-potential correlates of stimulus uncertainty. Science (80-) 150:1187–1188.

Tallon-Baudry C, Bertrand O (1999) Oscillatory gamma activity in humans and its role in object representation. Trends Cogn Sci 3:151–162.

Tallon-Baudry C, Bertrand O, Delpuech C, Permier J (1997) Oscillatory gamma-band (30-70 Hz) activity induced by a visual search task in humans. J Neurosci 17:722–734.

Thibodeau R, Aronson E (1992) Taking a Closer Look: Reasserting the Role of the Self-Concept in Dissonance Theory. Personal Soc Psychol Bull 18:591–602.

Tononi G (2004) An information integration theory of consciousness. BMC Neurosci 5:42.

Tononi G (2008) Consciousness as integrated information: a provisional manifesto. Biol Bull 215:216–242.

Tononi G, Edelman GM (1998) Consciousness and Complexity. Science (80-) 282:1846–1851.

Tranel D, Damasio H, Damasio AR (1995) Double dissociation between overt and covert face recognition. J Cogn Neurosci 7:425–432.

Turner M, Coltheart M (2010) Confabulation and delusion: a common monitoring framework. Cogn Neuropsychiatry 15:346–376.

Tusche A, Bode S, Haynes J (2010) Neural responses to unattended products predict later consumer choices. J Neurosci 30:8024–8031.

Tusche A, Kahnt T, Wisniewski D, Haynes J (2013) Automatic processing of political preferences in the human brain. Neuroimage 72:174–182.

Van Den Bussche E, Segers G, Reynvoet B (2008) Conscious and unconscious proportion effects in masked priming. Conscious Cogn 17:1345–1358.

Van Gaal S, de Lange FP, Cohen MX (2012) The role of consciousness in cognitive control and decision making. Front Hum Neurosci 6:121.

Van Gaal S, Lamme V a F, Ridderinkhof KR (2010) Unconsciously triggered conflict adaptation. PLoS One 5:e11508.

Van Gaal S, Ridderinkhof KR, Fahrenfort JJ, Scholte HS, Lamme V a F (2008) Frontal cortex mediates unconsciously triggered inhibitory control. J Neurosci 28:8053–8062.

Van Gaal S, Ridderinkhof KR, van den Wildenberg WPM, Lamme V a F (2009) Dissociating consciousness from inhibitory control: evidence for unconsciously triggered response inhibition in the stop-signal task. J Exp Psychol Hum Percept Perform 35:1129–1139.

Van Veen V, Carter CS (2002) The Timing of Action-Monitoring Processes in the Anterior Cingulate Cortex. J Cogn Neurosci 14:593–602.

Van Veen V, Krug MK, Schooler JW, Carter CS (2009) Neural activity predicts attitude change in cognitive dissonance. Nat Neurosci 12:1469–1474.

Vinck M, van Wingerden M, Womelsdorf T, Fries P, Pennartz CMA a (2010) The pairwise phase consistency: A bias-free measure of rhythmic neuronal synchronization. Neuroimage 51:112–122.

Vinckier F, Gaillard R, Palminteri S, Rigoux L, Salvador A, Fornito A, Adapa R, Krebs M-O, Pessiglione M, Fletcher PC (2015) Confidence and psychosis: a neuro-computational account of contingency learning disruption by NMDA blockade.

Wacongne C, Changeux J-P, Dehaene S (2012) A neuronal model of predictive coding accounting for the mismatch negativity. J Neurosci 32:3665–3678.

Wacongne C, Labyt E, van Wassenhove V, Bekinschtein T, Naccache L, Dehaene S (2011) Evidence for a hierarchy of predictions and prediction errors in human cortex. Proc Natl Acad Sci 108:20754–20759.

Wahlen A, Nahum L, Gabriel D, Schnider A (2011) Fake or fantasy: rapid dissociation between strategic content monitoring and reality filtering in human memory. Cereb cortex 21:2589–2598.

Wheatley T, Milleville SC, Martin A (2007) Understanding animate agents: distinct roles for the social network and mirror system. Psychol Sci 18:469–474.

White RW (1959) Motivation reconsidered: The concept of competence. Psychol Rev 66:297–333.

Whitson J a, Galinsky AD (2008) Lacking control increases illusory pattern perception. Science 322:115–117.

Winkler I, Karmos G, Naatanen R (1996) Adaptive modeling of the unattended acoustic environment reflected in the mismatch negativity event-related potential. 742:239–252.

Winkler I, Reinikainen K, Näätänen R (1993) Event-related brain potentials reflect traces of echoic memory in humans. Percept Psychophys 53:443–449.

Wolford G, Miller M, Gazzaniga MS (2000) The left hemisphere's role in hypothesis formation. J Neurosci 20:1–4.

Yabe H, Tervaniemi M, Reinikainen K, Näätänen R (1997) Temporal window of integration revealed by MMN to sound omission. Neuroreport 8:1971–1974.

Yellott JI (1969) Probability learning with noncontingent success. J Math Psychol 6:541–575.

Young AW (1998) Face and Mind. Oxford University Press.

Zaehle T, Bauch EM, Hinrichs H, Schmitt FC, Voges J, Heinze H-J, Bunzeck N (2013) Nucleus accumbens activity dissociates different forms of salience: evidence from human intracranial recordings. J Neurosci 33:8764–8771.