THÈSE

PRÉSENTÉE A

L'UNIVERSITÉ PIERRE ET MARIE CURIE

ÉCOLE DOCTORALE : Cerveau, Cognition, Comportement

Par Lucie CHARLES

POUR OBTENIR LE GRADE DE

DOCTEUR

SPÉCIALITÉ : Neurosciences Cognitive

MÉCANISMES CONSCIENTS ET NON-CONSCIENTS DE LA DECISION ET DE LA "META-DECISION"

Directeur de recherche : Stanislas DEHAENE

Soutenue le : 19 Septembre 2013

Devant la commission d'examen formée de :

M. Patrick CAVANAGH	LPP	Rapporteur
M. Nick YEUNG	Oxford University	Rapporteur
M. Lionel NACCACHE	UPMC	Examinateur
M. Sid KOUIDER	ENS	Examinateur
M. Mathias PESSIGLIONE	ICM	Examinateur
M. Stanislas DEHAENE	Collège de France	Examinateur

Acknowledgments

When I started my PhD four years ago, this page of acknowledgements was one of the first things I thought of writing as I already felt indebted to so many people who helped me start this adventure. It is therefore with great pleasure today that I express my warmest feelings of gratitude to all the people who contributed to make this time a rich, exciting and enlightening experience.

First and foremost, I would like to thank Stanislas Dehaene for giving me the great opportunity to work in Neurospin and for making these four years of PhD such a beneficial time. Beyond providing the most favorable environment to perform research, he has dispensed the appropriate dose of support and challenge to my work. His enthusiasm, his curiousity and his intellectual rigour will always be a model of what an excellent scientist ought to be.

My special thanks go to the Unicog lab and all of its past and present members, for having made scientific work such a fun occupation during these four years. I trust that it will stay, for many years to come, the welcoming and stimulating environment that I found it when I first arrived.

It has been a great honor and pleasure to be part of the consciousness group and I would like to thank all of its member, especially Filip Van Opstal, Jean-Rémi King, Moti Salti, Simon Van Gaal, Aaron Schurger and Jacobo Sitt for many fruitful discussions and laughs in the research of consciousness.

I am particularly grateful for the assistance given by the MEG team, in particular Virginie Van Wassenhove, Etienne Labyt, Marco Buiatti, Leila Rogeau and all MEG users.

I would like to thank the nurses Véronique Jolly-Testault, Laurence Laurier and Gaëlle Mediouni as well as the neurospin doctors Ghislaine Dehaene-Lambertz, Andreas Kleinschmidt, Caroline Huron, Josselin Houenou and Lucie Hertz-Pannier for their support in participant recruitment and testing. I am also very grateful for the assistance given by members of Saint-Anne hospital, Raphaël Gaillard, Isabelle Amado and Narjes Bendjemaa who made the testing of schizophrenic patients in Neurospin possible.

I must also give special thanks to all the people who were patient enough to listen and try to answer my never ending series of questions: Lauri Parkkonen, Alexandre Gramfort, Aaron Schurger, Christophe Pallier, Gaël Varoquaux, Fabian Pedregosa, Evelyn Eger and anyone else that I ever bothered with the phrase "Est-ce que je peux te déranger deux minutes ?".

I am committed to offer my thanks all the people who gave me a drive back to Paris late at night or who helped me to avoid waiting hours for the Saclay bus. As this list would take several pages to put down, I will only name Edith Le Floch whose car has always been a joyful and welcoming place to meet new neurospin fellows.

I would also like to say a great big thank you to all the individuals of the Homo Openspacius species, extinct or alive: Catherine Wacongne, Baptiste Gauthier, Flore Baronnet, Ghislain Bosquillon de Frescheville, Clio Coste, Lucille Lecoutre, Nicolas Zilber, Gabriel Garcia and honorary member Elodie Cauvet. This thesis would not have been the same without our never-ending debates on science, society, the geek way-of-life, our share of procrastination techniques and our laughs on stupid cat videos.

Working in Neurospin has given me the opportunity to meet amazing people, some of whom have

become life-long friends. Sebastien, in addition to being an always cheerful and stimulating colleague, your support in the highs and lows of the last four years has been invaluable. And I owe you, your Peugeot 206 and Neurospin parking lot my driving license. Anne, I will always be grateful that our PhD studies collided. Thanks for all the laughs, gossips and kindness. My phd could not have extended that long without you as an office-mate. And I will not forget the third member of the band: Diana, thanks for so many laughs and joyful moments, and for your very kind help in correcting this manuscript.

Last but not least, I wish to thank my parents for their wonderful help, love and support. Lionel, in 2006 you offered me the great book "Le Nouvel Inconscient" written by another Lionel, and without your gentle encouragement, stimulation and guidance, I would not have been able to complete this thesis. Anne, your kindness, your patience and your strength is a constant source of inspiration. My gratitude towards you both goes far beyond words.

Contents

I Review of the litterature **1**

1 Metacognition and Consciousness **5**
1.1 What is consciousness? . . . 5
1.1.1 The emergence of the notion of consciousness . . . 5
1.1.2 Paradigms for the scientific study of consciousness . . . 7
Pattern Masking. . . . 8
Metacontrast. . . . 8
Bistable perception. . . . 9
Binocular rivalry and continuous flash suppression. . . . 9
Crowding. . . . 10
Attentional blink. . . . 10
1.1.3 Signal detection theory and measures of consciousness . . . 11
1.1.4 The depth of non-conscious processing. . . . 17
1.1.5 Beyond visual awareness . . . 19
1.2 What is metacognition ? . . . 21
1.2.1 A few definitions. . . . 21
1.2.2 The feeling of knowing, the first research on metacognition . . . 22
1.2.3 Empirical approaches to measure metacognition . . . 23
1.2.4 The neuronal substrates of confidence judgements . . . 27
1.3 Models of confidence and error-detection . . . 29
1.3.1 Signal Detection Theory and confidence judgment . . . 30
1.3.2 Meta-d' . . . 33
1.3.3 Models of accumulation of evidence for first- and second-order decisions . . . 35
1.3.4 Dynamical models of error correction . . . 35
1.3.5 Alternative models of confidence judgments . . . 38

2 Error-detection, a simple metacognitive task **41**
2.1 A brief review on error detection . . . 41
2.2 The Error-Related Negativity: a cerebral marker of error detection . . . 44
2.2.1 Factors influencing the ERN amplitude . . . 45
2.2.2 Location of the origin of the ERN . . . 46
2.2.3 Functional Role of the ERN . . . 48
2.3 Consciousness and the ERN . . . 52
2.3.1 Variation of the ERN with confidence ratings . . . 52
2.3.2 The ERN in anti-saccade paradigms. . . . 55
2.3.3 An ERN for undetected errors . . . 55
2.3.4 An ERN in subliminal condition? . . . 56
2.4 Schizophrenia, Metacognition and Consciousness . . . 58
2.4.1 Psychopathology and the ERN . . . 58
2.4.2 Error detection and the ERN in schizophrenia . . . 61
2.4.3 Schizophrenia and Consciousness . . . 63

3 An experimental approach to study consciousness and metacognition 67
3.1 Masking study . . . 68
3.2 M/EEG, a powerful tool to study brain activity . . . 70
3.2.1 A brief description of MEG and EEG techniques . . . 70
3.2.2 The sources of electro-magnetic brain signal . . . 72
3.2.3 Reconstructing the source of M/EEG signal . . . 72
3.2.4 Why use simultaneous MEG/EEG recordings . . . 74
3.3 Decoding . . . 75
3.3.1 Mulitvariate Pattern Analysis . . . 76
3.3.2 Support Vector Machine . . . 77
3.3.3 Evaluating classification score . . . 79
3.3.4 The benefit and confounds of decoding . . . 80
3.4 Followed plan . . . 82

II Experimental contributions 85

4 Article 1 : Distinct brain mechanisms for conscious and subliminal error detection 87
4.1 Introduction to the article . . . 87
4.1.1 Context and goal of the study . . . 87
4.1.2 Experiment . . . 88
4.1.3 Summary of the results . . . 88
4.2 Article . . . 89

5 Article 2 : Decoding the dynamics of action, intention, and error-detection for conscious and subliminal stimuli 123
5.1 Introduction to the article . . . 123
5.1.1 Context and goal of the study . . . 123
5.1.2 Experiment . . . 123
5.1.3 Summary of the results . . . 126
5.2 Article . . . 126

6 Article 3 : Preserved unconscious metacognition and impaired conscious error-detection in schizophrenia 169
6.1 Introduction to the article . . . 169
6.1.1 Context and goal of the study . . . 169
6.1.2 Summary of the results . . . 170
6.2 Article . . . 171

III General discussion 219

7 Implications for the models of consciousness 223
7.1 The depths of non-conscious processes revisited . . . 223
7.2 Crossing of the threshold for conscious access : an all-or-none phenomenon ? . . . 226
7.3 Implication for the measure of consciousness . . . 229

8 Models of error-detection **231**
8.1 Computational models of the ERN 231
8.2 Dual versus single route model for decisions 233
8.3 Are confidence judgments and error-detection processes the same? 236

9 Perspectives **239**
9.1 Action and Perception: the same status for consciousness? 239
9.2 Metacognitive judgment of confidence outside of awareness 240

10 Conclusion **243**

Bibliography **245**

Part I

Review of the litterature

Summary of Part I

In this section, I will briefly discuss previous research on consciousness, describing how it emerged from the minds of philosophers and psychologists as an object that can be studied using the scientific method. I will then present the tools currently available to study consciousness, specifically experimental paradigms used to make a stimulus non-conscious and the measures used to assess consciousness. I will then highlight a few studies on non-conscious processing that have proved particularly influential in the current field of research, in particular those that went further than the question of perceptual awareness. Following this, I will briefly summarize the research on metacognition, from the first studies that raised the question of introspection to the recent methods developed to assess metacognitive knowledge, and then present evidence concerning the neural basis of metacognitive judgments. In my attempt to bring together these two research topics, consciousness and metacognition, I will discuss different theoretical models of decision and meta-decision and elaborate on how they relate to the question of consciousness and introspection.

In the following section, I will focus on the problem of error detection and demonstrate how it constitutes a relevant way to address the relationship between metacognition and consciousness. I will briefly show how the simple metacognitive task of error-detection corresponds to a known electro-physiological brain response: the error-related negativity (ERN). After discussing the factors influencing the ERN, its neuronal source, as well as its functional role, I will present existing research on this electrophysiological brain response and consciousness. I will discuss evidence concerning the variation of the ERN on subjective reports of confidence, and more specifically introduce how the ERN relates to the question of error awareness.

In the last part of this introduction, I will focus more specifically on the methods that are used in the current work to study the relation between consciousness and metacognition: in particular I will outline the paradigm, the brain imaging techniques and the analysis tools employed. I will then briefly present the work plan that has been chosen for this thesis.

CHAPTER 1

Metacognition and Consciousness

1.1 What is consciousness?

1.1.1 The emergence of the notion of consciousness

In 1637, René Descartes wrote the "Discours de la Méthode" in which he interrogated reasoning and the search for truth in science. This text, which constitutes one of the most fundamental contributions to the history of philosophy, describes the thought experiment of complete and systematic doubt, leading Descartes to identify the only truth that prevails, the existence of himself thinking. In the well-known sentence "Cogito ergo sum", "I think therefore I am", Descartes states that, regardless of the knowledge or the veracity of our reasoning, we cannot ignore our own experience of thinking. For Descartes, this notion constitutes proof that the mind dissociates from the body and constitutes an immaterial entity, justifying a dualist point of view known as the mind-body problem. Today, this question remains central. What causes our conscious experience? How does consciousness relate to the material world? Can we study consciousness simply as one feature among others of our cognitive system?

In the beginning of the twentieth century, almost three hundred years later, the study of consciousness has remained highly problematic. Following the emergence of psychology as a research discipline and the proposal by William James that consciousness should constitute the centre of the study of the mind, behaviourism came along to oppose this notion and proposed that on the contrary psychology ought to explain only objective facts. According to the behaviourist view, psychology should focus on observable behaviours of human and animals, without making assumptions about the activity of the mind, which could not be observed. Moreover, scientists were convinced that behavior not only ought to be studied without considering any underlying brain processes or abstract notions of thoughts or beliefs but also that it constituted the best evidence available to understand cognition. The introspective method was rejected by behaviorists like J. B. Watson or B. F. Skinner and questioning about conscious experience and the subjective mind was largely ignored.

However, the picture changed again in the middle of the twentieth century when cognitive sciences emerged and laid the foundations of modern psychology. For the first time, philosophers addressed the question of how one could possibly study consciousness scientifically, i.e. how the conscious thoughts and experience we have of our own mental life translate into brain activity and what their functions are. Philosophers played a key-role in this journey towards a scientific approach to the study of consciousness as they proposed a modern perspective on this notion. Fundamental questions were addressed that

continue to constitute the motivation and background of the scientific study of consciousness today.

The emergence of the notion of qualia constitutes a crucial step in establishing the key questions concerning consciousness as a field of research. Developed by Lewis (1929), the notion of qualia aimed to capture the richness of conscious experience and the subjective feeling that is associated with the life of the mind. "The quale is directly intuited, given, and is not the subject of any possible error because it is purely subjective." The notion of qualia encompasses the recognition of subjectivity as the unique gate to the world and the non-shareable aspect of our own private conscious experience. More importantly, it led philosophers to argue that two types of problems exist for consciousness: the hard problem and the easy problem. According to the philosopher David Chalmers, who developed this key distinction in the scientific study of consciousness (Chalmers, 1995), the core of the hard problem is to explain how and why we have qualia that constitute the nature of our phenomenal experience. In contrast, the easy problem consists of explaining how consciousness is linked to other cognitive functions such as integrating information, attending to an object and reporting our mental states. The easy problems are not easy in the sense that their solutions are easy to find. Rather, they are so-called because there is no doubt that they can be explained scientifically in terms of computational or neural mechanisms. In contrast, the hard problem must address the existence of subjective experience and does not seem to be solvable. This problem has been conceptualized by Thomas Nagel in his text "What is it like to be a bat?". Nagel (1974) develops the idea that a scientific explanation of consciousness in the sense of brain activation will omit an essential component of consciousness, which is what it feels like to consciously experience something. Therefore, there is an explanatory gap between materialist approaches of consciousness and the notion of qualia that cannot be overcome (Levine, 1983). At the same time, this argument also states that the cerebral basis of consciousness can be addressed by a scientific approach.

Interestingly, philosopher Daniel Dennett provocatively put forward the idea that the notion of qualia is invalid (Dennett, 1993). Indeed, he reproached the definition of qualia for being vague, and criticised the fact that it is either not usable or raises questions that are by nature not answerable. More importantly, he questioned the idea that qualia are more "special" than other properties of the world (Dennett, 1988). Dennett proposed an analogy to understand the lack of validity of the notion of qualia and the hard problem. He considered someone stating the following: *"That's all very well, all that stuff about DNA and proteins and such, but I can just imagine discovering an entity that looked and acted just like a cat, right down to the blood in its veins and DNA in its cells, but was not really alive."* (Dennett, 1993). Dennett suggested that the same argument could be made about the hard problem and qualia: the fact that we are not able to imagine how to solve the hard problem does not constitute, by itself, a justification of its existence. Dennett concluded *"I trust that no one thinks this is a good argument for vitalism. [...] The only thing this argument shows is that you can ignore all that and cling to a conviction if you're determined to do so"*. Therefore for Dennett, qualia and the question of the hard problem of consciousness can safely be ignored.

Interestingly, John Searle proposed a similar argument without refuting the existence of qualia

(Searle, 1998). According to him, qualia do exist. Moreover, qualia are the essence of consciousness and the question of consciousness cannot be separated from qualia: *The problem of consciousness is identical to the problem of qualia, because conscious states are qualitative states right down to the ground* (Searle, 1998). However, as stated by Dennett, the fact that we currently see the hard problem as a philosophical unsolvable question does not mean it will continue to be so in the future. In particular, Searle reminds us that there is no doubt that qualia are caused by brain activity. Therefore, the scientific study of consciousness should explain qualia. More importantly, Searle proposes that "the sense of mystery" that remains in consciousness will disappear once we have a precise theoretical and empirical account of conscious experience. Indeed, history of science is full of seemingly unsolvable problems that were eventually solved. Even in neuroscience, Searle notes: *"To Descartes and the Cartesians, it seemed mysterious that a physical impact on our bodies should cause a sensation in our souls. But we have no trouble in sensing the necessity of pain given certain sorts of impacts on our bodies. We do not think it at all mysterious that the man whose foot is caught in the punch press is suffering terrible pain. We have moved the sense of mystery inside. It now seems mysterious to us that neuron firings in the thalamus should cause sensations of pain. And I am suggesting that a thorough neurobiological account of exactly how and why it happens would remove this sense of mystery."*

While the debate remains lively, all philosophers agree that only a better understanding of the architecture and the neural substrate of our cognitive system will bring answers to the question of consciousness. In the present work, we propose to stay on this optimistic note, focusing on understanding the cerebral basis of consciousness and its relation to other cognitive functions, in particular metacognition.

1.1.2 Paradigms for the scientific study of consciousness

The adoption of consciousness as a scientific object of study implicated the development of operational measures of conscious state as well as paradigms for inducing conscious and non-conscious perception. These experimental conditions have been achieved in many different ways and have led to considerable debate. The contrastive approach of conscious versus non-conscious conditions as proposed by B. Baars is central to contemporary research on consciousness. The idea of this approach is to contrast the behavioural and cerebral response to a stimulus that is rendered unreachable to consciousness with the response to a consciously perceived stimulus (Baars, 1994). Typically, in this kind of paradigm subjects are asked to report their subjective feeling of visibility while images are flashed either consciously above the threshold for reportability or subliminally below the threshold for consciousness. We can study the brain responses to these images, in particular when they are not perceived, following the idea that the specificity of conscious processes will emerge by contrasting the behavior and brain responses to images presented consciously from those to images presented non-consciously. Many types of techniques can be used to render an image subliminal and various paradigms have been developed. In this respect, masking paradigms probably constitute the simplest method. In this type of paradigm, a target stimulus is presented very briefly (usually between 10 and 30 ms) and is immediately followed by

another stimulus, the mask, that is usually more salient and clearly visible. The target stimulus is very easily not perceived by the subject, especially when the delay between the onset of the target and the onset of the mask (the so-called Stimulus Onset Asynchrony, SOA) is very short. In this case, the subject only consciously perceives the mask. However when we increase the delay slightly, for example above 50 ms the target stimulus gradually becomes more visible, the subject easily perceiving the succession of the two stimuli.

Many variants of this type of paradigm exist relying on different mechanisms to create the masking effect. Typically, two main types of masking are documented according to the spatial relationship that exists between the target and the mask (Enns and Di Lollo, 2000).

Pattern Masking. The first type of masking consists of presenting the mask at the same location as the target stimulus, with the mask superimposing on the target, in a manipulation referred to as "pattern masking". This method, which can very efficiently reduce the visibility of the stimulus to complete invisibility, has often been used with the association of two masks, one preceding and one following the target stimulus. This technique commonly referred to as "sandwich making" relies on the presentation of a long backward mask followed by a rapid target presentation and a short forward mask (Kouider and Dehaene, 2007). Stronger masking is obtained when the mask is made of scrambled images similar to those of the target stimulus (see Figure 1.1). While the exact mechanisms of this type of masking remain poorly understood, it is thought that two types of effects may be involved (Enns and Di Lollo, 2000). First, the fact that the mask is presented at the same location and almost at the same time as the target stimulus could result in the merging of the two stimuli into a single noisy pattern in the early stage of visual processing, creating a phenomenon of "integration masking". Second, it has been proposed that at higher stages, competition for higher-level computational resources between the mask and the target could produce a form of "interruption masking", blocking the processing of the target at an early stage of visual processing. Interestingly this form of masking permits a relatively longer presentation of the target stimulus compared to other techniques.

Metacontrast. A second type of masking that has been highly employed and thoroughly documented in many experiments is "metacontrast masking". Here, the mask stimulus appears at a close adjacent location to the target but in a non-overlapping manner with the contours of the mask matching the contours of the target. Interestingly, metacontrast masking has a slightly different timing to classic pattern masking: visibility of the target stimulus remains unimpaired for very short or very long SOAs while intermediates SOAs are characterized by decreased visibility and reduced objective performance (Breitmeyer and Ogmen, 2006). Several alternative theories have been proposed to explain the mechanisms of metacontrast masking. In particular it has been suggested that metacontrast could result from the interaction of two distinct pathways for visual perception, exhibiting different characteristics to spatial frequencies (Bruchmann et al., 2010). According to this view, the response to the target by a sustained pathway would be suppressed by the transient response evoked by the mask. Interestingly, these two

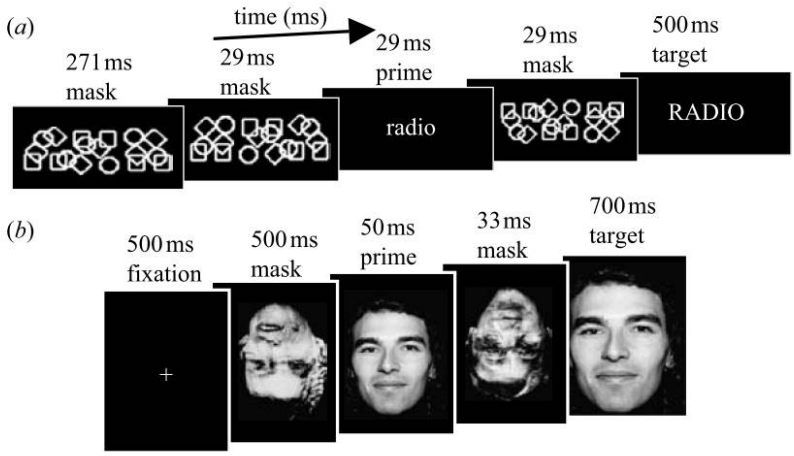


Figure 1.1: Example of pattern masking for words and faces (from Kouider and Dehaene, 2007). The two panels present experimental paradigms based on repetition priming in which the prime stimulus is invisible. (a) In visual word repetition priming, a prime word is briefly flashed, preceded and followed by two masks, before the target word is presented in a different case. If the two words are identical, the prime word will facilitate the processing of the second word, reducing overall response-times. (b) In the face repetition priming, the prime can be the same person or a different person to the target face. Importantly, masks are made of parts of reversed faces, which increase the efficiency of the masks. To avoid simple superimposition effect, prime size is reduced by 80% compared to the target. Again, repetition of the same face will induce priming effect on response times.

pathways have been linked to the magno- and parvocellular pathways of the visual system, suggesting a low-level mechanistic explanation of this form of masking.

Bistable perception. Other alternative techniques have been developed in order to render a stimulus non-visible. In particular, some stimuli can be perfectly perceived by the eyes and the visual cortex while their meaning or their identity is not recognized consciously. Indeed, such effects have been used by artists for a long time, before being used experimentally. One example of such a manipulation is bistable perception. Bistable stimuli are characterized by the possibility of being interpreted as two different objects. A very well known example is Necker's cube (Figure 1.2) in which a three-dimensional cube is plotted onto a two-dimensional space, leaving open the possibility to see the cube as being oriented toward or away from the viewer. Crucially, conscious perception of the cube is always dominated by one interpretation, reflecting the competition between the two percepts to reach consciousness. This type of paradigm is particularly interesting as it allows the study of the effect of conscious perception alone, while the stimulus display is kept strictly constant.

Binocular rivalry and continuous flash suppression. In a similar vein, but on a lower perceptual level, binocular rivalry permits the presentation of a stimulus for several seconds, without it ever reaching consciousness. In this case, a different stimulus is presented separately to each eye of the viewer. If the two displays are clearly distinct, they are not fused by the brain but on the contrary are perceived

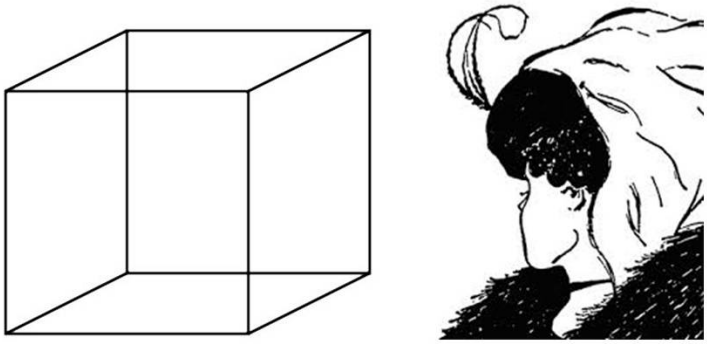


Figure 1.2: Examples of bistable images. While Necker's cube (on the left) is ambiguous in terms of its spatial orientation, pointing either forward or backward from the paper sheet, the second image can be seen as an old woman's profile or as a young woman looking away. At each instant, one interpretation overrides the other, creating a bistable state of perception.

alternatively, each image competing to dominate conscious visual perception. Importantly, it is possible to manipulate which image of the two eyes is going to be consciously seen by the subject. This is achieved for example using the continuous flash suppression technique (CFS). In this type of paradigm, a constant stimulus is presented to one of the two eyes while a series of rapidly changing stimulus such as Mondrian patterns are presented to the other eye (Tsuchiya and Koch, 2005). The result of such a manipulation is that the static image is suppressed and only the changing stream of images is perceived, with the effect able to last several seconds.

Crowding. Other forms of masking can take place when a task is performed on objects situated in the periphery of the visual field. Masking by crowding, for example, is observed when an object in the periphery (such as a faint dot) is masked by a neighboring object which while also in the periphery is more salient (such as a written word). In this case, a subject who is fixating on the centre of the screen fails to report the presence of the dot, while consciously perceiving the written word. Interestingly, it has been proposed that crowding not only results from the poor resolution of visual or attentional mechanisms in the periphery but may also constitute an artifact of the preparation of eye-movements: when shifting spatial attention from the fovea to the periphery, automatic triggering of saccade mechanisms, in particular image displacement, might bias image orientation statistics creating the phenomenon of crowding (Nandy and Tjan, 2012). This paradigm is particularly interesting as it allows one to study conscious perception while a stimulation is kept constant.

Attentional blink. Following the same line of research, other paradigms can be used to make a stimulus undetectable to the subject by directly manipulating attentional mechanisms. While almost all masking techniques partially rely on attentional mechanisms to create the condition of invisibility, some

of them rely almost solely on this aspect and prove very useful in creating a complete absence of conscious percept. These paradigms constitute an important field of research and are the subject of very lively debate in the scientific community, as the link between attention and consciousness constitutes a key point in understanding the architecture of our cognitive system. Not all of the paradigms manipulating attention can be presented in this thesis. However, one type of non-conscious perception, thought to rely on attentional processes, which should be noted is the attentional blink (AB). The attentional blink is observed in rapid serial visual presentations in which a continuous stream of graphical objects such as letters or numbers is presented to the subject centrally (Luck et al., 1996; Marois et al., 2000). When the subject is instructed to perform a task on two of the objects displayed consecutively, he or she will often accurately perform the task on the first object while missing the second object. Interestingly, this effect occurs only when the two objects are not presented successively (one right after the other) but have at least one display between them, a phenomenon called lag-1 sparing. It has been proposed that the AB occurs as a result of the competition between the target stimuli to access a central stage process that acts as a bottleneck for the processing of the two stimuli (Marti et al., 2012; Zylberberg et al., 2010). While further evidence is needed before conclusions regarding the validity of this model can be drawn, AB, as with other paradigms manipulating attention, provides evidence on which processes can occur non-consciously and help to gain a global view of the architecture of our cognitive system that allows consciousness to develop.

1.1.3 Signal detection theory and measures of consciousness

Regardless of the paradigm used to achieve non-conscious stimuli, a key question that remains is how to assess the subject's conscious experience. What is a good measure of consciousness? We have seen that the study of consciousness is not dissociable from subjectivity, the object studied being the experience of conscious perception itself. Therefore, introspective report of the subject seems to be the key measure to assess conscious experience. Following this idea, the modern study of consciousness by the cognitive sciences reused introspective methods developed by psychologists during the nineteenth century. However, instead of complex and descriptive verbal reports, as developed by psychologists such as W. M. Wundt, which proved difficult to replicate and analyze, modern psychology uses mostly categorical choice to estimate perception. In the case of visual awareness for example, the most common measure used is subjective report of visibility of the stimulus, the subject performing a binary choice between *seen* and *unseen* responses according to his or her perception of the stimulus.

Looking at the question from this angle however, the study of consciousness can appear unrealistically simple. Indeed, several immediate critiques can be made to this approach:

1. Can the experimenter trust subjective reports?
2. Are different subjects reporting their perception identically?
3. Are some subjects more conservative or more liberal in their perceptual judgment?

4. Is subjective perception always binary, a stimulus being only consciously perceived or completely missed?

5. When reporting seeing something, can we systematically access the identity of this object or do we simply detect its presence?

These critiques have led to many discussions and debates on how to properly assess consciousness and different methods have been proposed to circumvent potential pitfalls.

Setting aside questions 1 and 2, which are concerned with the veracity of subjective report, we will address in the present paragraph the issue of bias in responding as presented in question 3. This problem is particularly important in the case of detection tasks, when the subject is asked simply to detect the presence or the absence of a target masked stimulus. While, intuitively, reporting the presence of an object appears the most relevant approach to contrasting different conditions of visibility, there are, in fact, many confounds which can be induced in such a task. Most critical is the question of bias towards a response. When a perceptual judgment is uncertain but the subject is asked to produce a binary choice, some subjects might decide to be conservative and not hazard a "*seen*" or "Target present" response while others might adopt a liberal criterion and use these responses with less hesitation.

These questions are not unique to consciousness studies and have long been addressed by engineers and physicists. Indeed, during the Second World-War engineers developed a mathematical theory that accounted for the specificity and sensitivity of their radars. To detect enemy aircraft, the soldiers had to determine if the spots seen on radars screens were real planes, or simple noise (such as birds or random dots of light). The problem was that there were no simple criteria for making these kinds of decisions. Both choices had a risk attached: if an enemy went undetected, people could be killed whereas if noise was interpreted as an enemy, this false-alarm would result in loss of time and money. To provide a rational way to make these decisions, engineers developed Signal Detection Theory (SDT) which provides a framework to conceptualize and quantify the question of specificity versus sensitivity and to find the optimal decision threshold.

Let us consider a subject who has to detect the presence of a subliminal stimulus. Sometimes the stimulus is presented (for example in 50 % of the cases) and the subject has to report seeing it. However, sometimes the stimulus is absent, random noise replacing the stimulus and the subject still has to detect whether something was presented or not. If the stimulus is difficult to perceive, like the faint light on a radar screen or a very strongly masked subliminal stimulus, errors occur. From the reports of the subject, we can then create a contingency table (Figure 1.3) corresponding to the number of times the subject responded present when the target was actually present (hit), the number of times the subject responded absent when the target was indeed absent (correct reject), the number of times the subject responded absent when the target was nonetheless present (miss) and the number of times the subject responded present when the target was in fact absent. We can calculate the conditional probability according to this

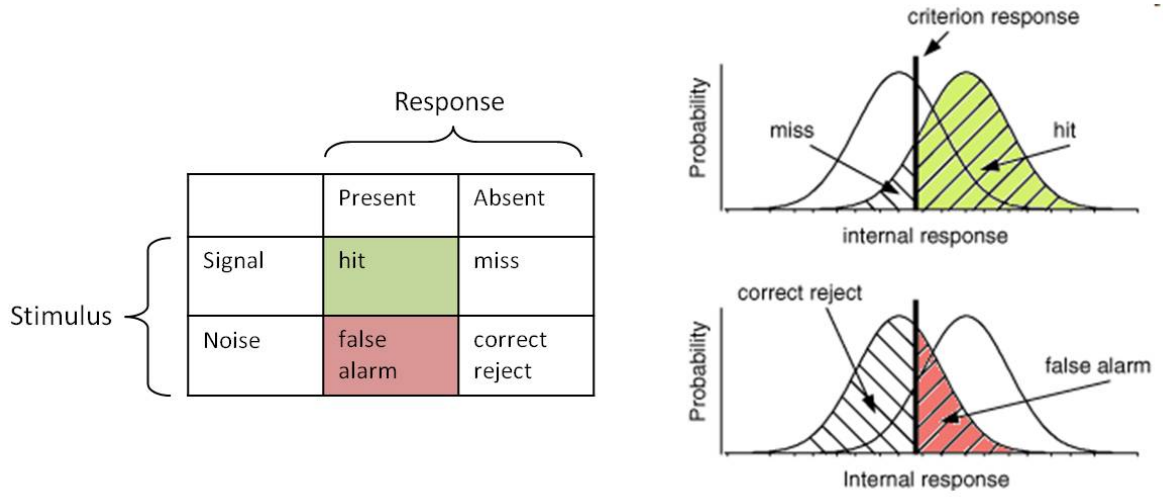


Figure 1.3: Contingency table and mathematical model for Signal Detection Theory. The contingency table depicts all possible types of trials when considering a simple detection task in which a noisy stimulus is either present or absent. Responses of the subjects allow the separation of trials according to the conditional probabilities of whether the subject saw the target stimulus when it was indeed present (hit) or on the contrary responded that the stimulus was present when it was in fact absent (false-alarm). Such decisions can be modeled as the internal response probability of occurrence when the stimulus is present (noise + stimulus) and when it is absent (noise only). Evidence for each trial falls on the decision axis and is compared to the position of the response criterion (vertical line), producing a response according to which side of the criterion it has fallen. This model allows the determination of the distribution of hits and false-alarm as shown by the area-under curve colored on each plot (adapted from D. Heeger, Department of Psychology, New York University)

contingency table, for each line of the table:

$$h = \frac{Numberofhit}{NumberofSignal\ trials} \tag{1.1}$$

and

$$f = \frac{Numberoffa}{NumberofNoise\ trials} \tag{1.2}$$

where h and f represent respectively the hit rate and the false alarm rate. As these values represent the conditional probabilities according to the signal presence, we can deduce from the other values of the table:

$$miss\ rate = 1 - h \tag{1.3}$$

and

$$correct\ reject\ rate = 1 - f \tag{1.4}$$

Intuitively, however, we can see that these values are redundant to the *h* and *f* values. Indeed, if you have information containing only *h* and *f*, it is clear that the behavior of the observer can be well characterized. For example, within the sentence *"When the target was present, Subject A responded in 90 % of the trials that he saw the target but when the target was absent he reported 80% of the trials that he saw it too"*, there is sufficient information to determine that subject A is not very precise and should not be trusted. Values of hit rate and false-alarm rate indeed reflect respectively the sensitivity ("how much can we detect") and the specificity ("how much what we detect reflects the true state of reality") in the response.

SDT proposes a statistical model to account for these values. The assumption behind it is that our detection system, as with any other detection system, electronic or biologic, is not perfect and carries some intrinsic random noise. Therefore the exact same stimulation will not always correspond internally to the same amount of evidence. Rather, it follows a Gaussian distribution in which the values corresponding to the presentation of the stimulus and the presentation of the noise falls around a mean value (see Figure 1.3): 0 when only noise is presented and *d'* when signal is presented in addition to noise. To make the decision to respond present or absent, SDT assumes that we set a criterion value on the decision axis, any responses falling on the left of this axis corresponding to a "stimulus absent" response and any response falling on the right leading to a "stimulus present" response.

This model of decision allows us to distinguish two different aspects of the decision process: the sensory process, corresponding to the perception of changes in physical stimulation along the decision axis, and the strategic process, corresponding to the bias in the decision, reflected in the criterion chosen. While the sensory process is characterized by the shape of the distributions, in particular how different are their mean values on the decision axis (*d'*) as well as the variance among each distribution and how much they overlap (Figure 1.4), the criterion *c* reflects the bias towards one response or the other. Importantly, the criterion *c* can be set optimally, at equal distance from the two distribution means, allowing the best performance considering the perceptual sensitivity to be obtained.

Setting these two values *c* and *d'*, we can understand how they relate to the previously seen hit and false-alarm rate (Figure 1.3). Assuming that the two distributions for noise and signal have equal variance, we can now estimate these values. Therefore:

$$c = -\frac{Z(h) + Z(f)}{2} \tag{1.5}$$

where Z is the inverse of the cumulative gaussian distribution. As Gaussians are symmetrical and $correct\ reject\ rate = 1 - f$, we arrive at the equation 1.5. Similarly,

$$d' = Z(h) - Z(f) \tag{1.6}$$

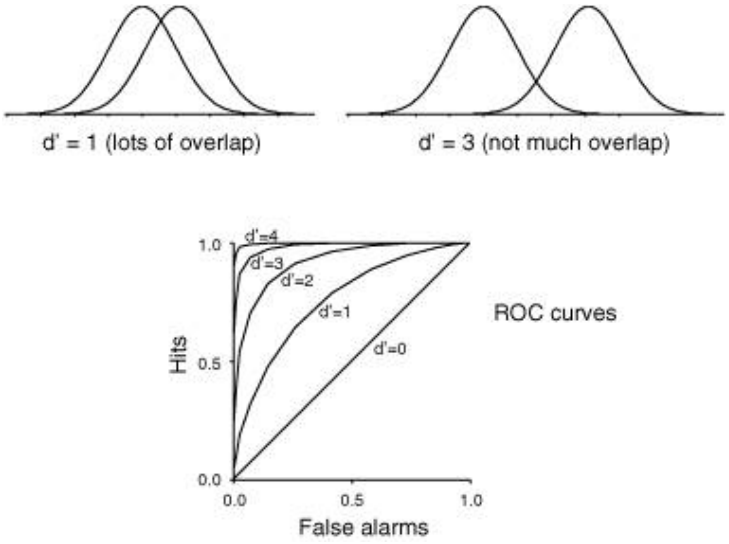


Figure 1.4: ROC curve and *d'* measure. The top graph represents the internal response probability for target present and target absent for different signal strengths, corresponding to two different d' values. The corresponding ROC curves are plotted below and correspond to the plots of hit-rate versus false-alarm rate when keeping *d'* value constant and varying response criterion. A diagonal ROC curve corresponds to a null *d'* and a total overlap of the two distributions (from D. Heeger, Department of Psychology, New York University).

Therefore, the computation of the *d'* value can provide an unbiased measure of the sensitivity to the masked target. As we have seen however, a given *d'* measure can be associated with a range of criterion or bias values. In other words, different pairs of hit and false alarm rates can correspond to the same *d'*. Figure 1.4 plots the values of h and f associated with the same *d'*. This curve called the Receiving-operator curve (ROC curve) captures in a single graph the various alternatives while keeping *d'* constant but moving the criterion to higher and lower levels. The area under-curve (AUC) of the ROC curve can be computed, also providing an unbiased measure of the detection sensitivity.

AUC and *d'* allow one to obtain an unbiased measure of detection sensitivity across subject. In particular, a *d'* value of 0 when detecting the presence of a masked target is considered characteristic of true subliminal conditions as the subject is completely unable to predict the occurrence of the target. While this analysis allows one to obtain a much clearer idea of the mechanisms leading to the decision, it also takes us far away from the problem of consciousness. In particular, can consciousness be characterized solely in terms of bias and sensitivity, putting completely aside the question of the subjective experience of the subject? Furthermore, while a very small *d'* can indicate that a subject is not able to detect the presence of the target and therefore is probably not consciously perceiving it, is a non-null *d'*automatically associated with conscious perception? Following this line of questioning, detection *d'* has been criticized when used as a measure of visual awareness. While *d'* allows one to assess the degree of detectability of the stimulus and, in doing so, provides a quantitative measure on how well a stimulus

is perceived, it seems ill equipped to characterize conscious experience and therefore cannot be regarded as an adequate measure of consciousness.

An interesting alternative to these objective measures of detection is to adopt the opposite approach and use a continuous subjective scale to assess visibility (Overgaard et al., 2010; Sandberg et al., 2010). This approach is directly relevant to our fourth question *"Is subjective perception always binary, a stimulus being only consciously perceived or completely missed?"*. Continuous scales of visibility involve asking the subject to rate the visibility of masked target on a scale with different possible levels, ranging from total to full visibility. For example, the perceptual awareness scale (PAS) uses four levels "No experience", "Brief glimpse", "Almost clear experience" and "Clear experience" to assess subjective visibility (Ramsø y and Overgaard, 2004). In theory, such a measure should allow the uncertainty of perception judgment to be captured by truly reflecting the perception of the subject. Importantly, they do not rely on the assumption that conscious perception is a binary phenomenon. However, an interesting study performed by Sergent and Dehaene (2004a) showed that even when provided with a continuous scale comprising up to 21 possible positions, subjects still use the scale in a binary manner (Sergent and Dehaene, 2004a). Indeed, in this experiment, participants used almost exclusively the extreme ends of the scale to report their visibility, intermediate levels being systematically ignored. This striking finding has been interpreted by the authors as the all-or-none characteristic of conscious perception, linked to a non-linear transition between non-conscious and conscious perception. In other words, conscious perception reflects the output of the decision, after transforming it to a binary judgment. While this hypothesis can be further discussed, the experimental findings nonetheless demonstrate the validity of binary responses in measuring consciousness.

A slightly different approach to the classic contrasting method is the priming technique. This method is particularly relevant when the question asked is not simply what the distinction is between conscious and non-conscious process but rather which process can operate non-consciously and how non-conscious information can influence conscious operations. In this kind of paradigm, a masked prime stimulus is presented, followed by a target stimulus on which the subject needs to perform a categorization task. In repetition priming, the prime may either be identical or different from the target. More generally, the prime is considered congruent or incongruent with the identity of the target, congruent primes facilitating the processing of the target, an effect noticeable on reaction-time (RT), accuracy and brain responses. Importantly, the invisibility of the prime stimulus needs to be addressed specifically, by subjective reports and usually by a forced-choice detection task, making necessary a prime-absent condition that allows the application of signal detection theory. Interestingly, while the prime needs to be relevant to the processing of the target stimulus for the priming effect to occur, it does not need to be directly relevant to the task at stake. For example, when presenting target stimuli corresponding to famous or unknown faces and the subject's task is to judge the faces' familiarity, a prime comprising an unrelated face with a negative or positive emotion would be irrelevant to the task, providing an independent measure of emotion priming without task-related artifacts. This effect constitutes an indirect measure of non-conscious information on conscious decision making, a particularly elegant method with which to

address the question of the depth of unconscious processes.

Finally, it has been proposed that visual awareness can be assessed by metacognitive accuracy or confidence judgments on the visibility response (Kolb and Braun, 1995; Rounis et al., 2010; Lau and Passingham, 2006). Intuitively, the ability to evaluate and judge our own mental process seems to be tightly linked to a type of reflexive, introspection process, typically linked to conscious experience. Several studies found that when performing a task in a heavily masked condition, subject could perform above-chance with no concurrent insight into their own ability to perform the task (Kolb and Braun, 1995; Lau and Passingham, 2006; Rounis et al., 2010; Szczepanowski and Pessoa, 2007; Weiskrantz, 1996), as evident in the lack of accurate subjective ratings on their own performance. In contrast, accurate metacognitive sensitivity implies that subjects are able to introspect their own cognitive process and retrieve information on their own mind, a function that seems tightly linked to conscious access (Kolb and Braun, 1995; Lau and Passingham, 2006; Rounis et al., 2010). However, the link between metacognition and consciousness remains to be tested in an objective manner, in the same way that first-order performance is analyzed by signal detection theory when performing a task on subliminal stimuli. This question will be investigated further in the following sections and in the discussion section of this thesis.

1.1.4 The depth of non-conscious processing.

As we have seen, many methods have been developed to render a stimulus subliminal as well as to measure how such stimuli influence behaviour and brain processes. Interestingly, research on the question of consciousness has been built not specifically on the questions of the role of consciousness, the specificity of conscious brain process or the nature of conscious perception but rather on the depth of non-conscious processing. Following S. Freud's proposal of a complex unconscious mind which comprises the origin of the majority of our behaviors, the depth of non-conscious processing became a popular question that has strongly impacted scientific research on consciousness. An important controversy worth mentioning on the subject occurred in the mid fifties. With the rise of film industry and optical effect on images, a study was conducted in which sentences such as "drink coca-cola" and "eat pop-corn" were flashed repeatedly as subliminal images during a film, remaining undetectable to the viewers. The man who organized this setting claimed that, as a result of the presentation of these subliminal images, the sales for these products increased. Following this discovery, subliminal images were banned from advertising and films in the US on the grounds that they could influence the behaviour of consumers. While the author of the study later admitted that his results were invented, the myth of subliminal images was nonetheless born.

Scientists have not been immune to this fascination for unconscious processes and this question has led to important scientific research on the depth of the "unconscious" that has greatly contributed to the scientific understanding of consciousness. In this respect, the role of neurology and neurophysiology has been very important, in particular with the study of patients suffering from specific lesions

impairing their conscious experience. The discovery of blindsight patients has been a crucial result in proving the existence of non-conscious perception. Blindsight describes a phenomenon in which patients with a specific lesion to the primary visual cortex retain the ability to discriminate and localize visual stimuli presented in their blind hemi-field, despite denying any conscious experience of the stimuli (Weiskrantz, 1986; Azzopardi and Cowey, 1998; Weiskrantz, 1996). GY, perhaps the most famous blindsight patient to have suffered from a lesion in the right visual cortex from an early age, shows an impressive ability to detect objects placed in his lesioned hemi-field, while claiming a total absence of conscious experience. This patient can, for example, detect the position of a stimulus presented briefly at different eccentricities in the cortically blind field and perform orientation or motion discrimination tasks. Similarly, when asked in a forced-choice manner, blindsight patients can perform appropriate actions to avoid or manipulate objects that they deny seeing. Other types of patients show interesting deficits regarding consciousness. Hemi-neglect patients, for example, suffer from parietal lesions often following a stroke. These lesions provoke attentional deficits leading the patients to ignore the contralesional side of the visual field, as if one side of sensory space was non-existent. Such patients will for example not eat the food on one half of their plate or produce incomplete drawings, representing only one side of the depicted objects. Interestingly, when presented images both in their impaired and in their preserved hemi-field, while being unable to identify the one in their blind hemi-field, their performance in discriminating the neglected image will exceed chance-level.

Having accumulated strong evidence for the existence of non-conscious processing in patients, many studies tried to investigate the depth of non-conscious processing in normal subjects. In a series of articles using the priming method, the non-conscious influence of subliminal primes was demonstrated. A founding article (Greenwald et al., 1996) showed that prime words influence semantic analysis of following target words. The authors showed that when classifying target words as pleasant or unpleasant, words preceded by a congruent prime showed improved performance, even when discrimination *d'* for the prime was at chance. Following this study, the influence of subliminal prime was shown to a greater extent, suggesting that not only could unconscious stimuli be processed up to semantic level but also that such an effect could be recorded in the brain with electro-physiologic and haemodynamic techniques (Luck et al., 1996). In an interesting study (Dehaene et al., 1998), subjects were asked to perform a number comparison task on a target number while a prime preceding the target was presented either consciously or non-consciously. Crucially, the prime was either congruent or incongruent with the number. In congruent conditions, subjects were faster to perform the number comparison task. More importantly, for the first time, this study showed that the effect of priming on reaction-times translated in detectable neural activity in the motor cortex when preparing the response to the target stimulus, recordable both with electroencephalography and functional-resonance imaging techniques. This work was replicated but with prime words that were never consciously presented to the viewer (Naccache and Dehaene, 2001). The results again showed an effect of priming, demonstrating that even for a novel stimulus, non-conscious processing up to the semantic level can modulate decision.

These results were further confirmed by several studies, showing a robust effect of semantic priming

(Weibel et al., 2013; Reynvoet and Ratinckx, 2004; Van den Bussche et al., 2009). Additionally, it was shown that non-conscious semantic content can modulate perception, improving visibility of the same words masked more strongly (Gaillard et al., 2006). Finally, it was shown that even when considering masked targets in isolation, non-conscious semantic activity can be evoked (Naccache et al., 2005) as revealed by a study in patients implanted with intracranial electrodes within the amygdala and showing specific response to subliminal emotional words. Indeed, fast processing of written words was confirmed by a more recent priming study showing that word recognition can operate non-consciously even for hand-written stimuli (Qiao et al., 2010). Although criticisms have been raised regarding the methodology of subliminal semantic priming (Kouider and Dupoux, 2004), especially when using word stimuli, these results were further confirmed by studies using different paradigms (Yeh et al., 2012) showing that semantic information can be integrated non-consciously.

Interestingly, subliminal semantic processing is not limited to written words. Subliminal digit primes were shown to influence decision, suggesting that numbers are processed in a complex manner and arithmetic operations are performed non-consciously (Van Opstal et al., 2011). More recently, Sklar et al. (2012) used continuous flash suppression to show that semantic violation of arithmetic operations can be detected even unconsciously. In a similar manner, a study investigated the neuronal basis of syntax and its relation to consciousness, showing that syntactically incorrect sentences evoked early brain responses, even when violations remained consciously undetected by the subjects (Batterink and Neville, 2013).

1.1.5 Beyond visual awareness

While visual awareness, in particular for words and number stimuli, has been widely investigated and proof of the existence of non-conscious perceptual processes has been found, the question of whether these results are limited to visual processing remains. Can they be replicated in other sensory modalities or extended to higher-order cognitive functions? In particular, some authors have looked for correlates of non-conscious auditory processes. Attempts to use priming in audition have proved somewhat more difficult than visual priming, as much less is known regarding how to mask auditory stimuli reliably. However, some studies managed to do so successfully (Kouider and Dupoux, 2005), using techniques in which the physical properties of the auditory stimulus are not degraded, but the target speech is hidden in a stream of non-speech sounds with similar spectral characteristics. These results have been replicated (Kouider et al., 2010; Dupoux et al., 2008; Davis et al., 2007) suggesting that non-conscious priming effect and subliminal processing (Sadaghiani et al., 2009) can occur in modalities other than vision.

However a question that remains unanswered is whether non-conscious processing can occur outside of sensory areas for higher cognitive functions. While reaching semantic processing already represents quite an important step in the hierarchy of stimulus processing, the question of whether other cognitive functions distinct from perceptual processes can be triggered non-consciously constitutes an important point to address. This question is also crucial because several models of non-conscious effect rely

strongly on sensory input to account for it. For example, it has been proposed that non-conscious perceptual information manages to enter the cognitive system as a feed-forward sweep, activating areas along its way (Lamme and Roelfsema, 2000), but progressively vanishing with the depth of processing.

While non-conscious processing of semantic category was demonstrated, the search for complex non-conscious processes has been pushed even further by studies on implicit learning. It was shown that subjects learn sequences of stimuli that are repeated in an implicit manner, reaction-time getting faster with learning and slowing-down for new sequences (Curran and Keele, 1993; Curran, 1995; Cohen et al., 1990; Reed and Johnson, 1994). Importantly, it was also shown that subjects remained unaware of the existence of repeated sequences, as evidenced by their failure to report them (Curran and Keele, 1993). However, such conclusions were debated, on the grounds that it was sometimes difficult to determine whether implicit learning was truly unconscious given that sequences of stimuli were always presented in a fully conscious manner. The use of specific methods to test whether learned information remained unbeknown to the subject was therefore proposed (Destrebecqz and Cleeremans, 2001). By means of an inclusion/exclusion paradigm whereby subjects are asked to exclude learned patterns from their responses, it was shown that subjects are unable to apply explicit rules on the learned stimuli, thus suggesting that they remained truly unconscious. More recently, it was shown that learned association of words presented subliminally were indeed encoded in brain activity and influenced further retrieval of the learned words (Reber et al., 2012), suggesting that learning mechanisms could indeed be triggered completely outside of consciousness.

Two important studies further confirmed that non-conscious processing can be extended to higher-order cognitive functions involved in learning and motivation (Pessiglione et al., 2008; Pessiglione et al., 2007). In a first study, the authors presented masked incentives (coin images) to subjects while they performed a hand-grip force task (Pessiglione et al., 2007). The exact level of motivation was manipulated by presenting either a large or a small incentive for the task. Simultaneously recording skin conductance, hand-grip force and brain activity, the authors showed that subliminal incentives modulated brain activity as well as behavioural responses,and thus showed the effect of subliminal motivational cues on behaviour. In a second study, the authors went even further to show that subliminal abstract stimuli arbitrarily associated with larger rewards were preferentially learned by the brain (Pessiglione et al., 2008), allowing learning and motivation system to operate completely outside of awareness. This results have been further confirmed by several studies(Schmidt et al., 2010; Capa et al., 2011) demonstrating that subliminal reward cues indeed modulate performance.

Similarly, several studies have investigated the link between consciousness and cognitive control, in particular response inhibition, extending non-conscious operations to not only slow learning processes but also to trial-by-trial control of behavior (Cohen et al., 2009; van Gaal et al., 2010; van Gaal et al., 2008; van Gaal et al., 2009). The authors showed that, not only do subjects slow down their responses when presented subliminal no-go signals but also that these non-conscious stop-signals modulate electrophysiological brain responses. In particular, both early responses such as the N2 and later events such as the P3 were affected, these components being linked to activation in prefrontal cortex (van Gaal et al.,

2008; van Gaal et al., 2009) and taken to reflect the triggering of the inhibition network. These findings were replicated and extended using fMRI (van Gaal et al., 2010), showing that inferior frontal cortex (IFC) and the pre-supplementary motor area (pre-SMA) are activated by non-conscious stop-signals.

Interestingly, other elements of cognitive control have been shown to be modulated non-consciously. In particular, several studies suggest that task-set preparation may be triggered in subliminal conditions (Lau and Passingham, 2007; De Pisapia et al., 2011; Reuss et al., 2011; Zhou and Davis, 2012; Mattler, 2003; Martens et al., 2011). Lau and Passingham (2007) used a priming method to evaluate if task-switching subliminal cues could influence behavior. Interestingly, they found that, when presented with subliminal primes coding for the alternative task, subjects were less accurate in performing the non-cued task and further activity in regions associated with the task decreased while activity in the region associated with the alternative task increased. It was further demonstrated that the cues did not need to be presented consciously to the subjects to observe priming of task-set, suggesting that the results could not be explained by low level perceptual effects (Zhou and Davis, 2012).

On a parallel line of research, several studies have investigated how consciousness may be linked to action and the sense of agency. The work of M. Jeannerod in this respect had a crucial impact on the field of consciousness research (Fourneret and Jeannerod, 1998; Jeannerod, 2003). To test whether we possess good insight into our motor actions, Fourneret and Jeannerod developed experiments in which subjects had to draw a straight line without seeing their actual hand during the motor action but only a computer screen feedback. Crucially, the authors biased the visual feedback given to the subjects and investigated their perception of the movements. Interestingly, while subjects corrected their actions online, taking into account the experimental bias, they nonetheless failed to report the deviation of their own movement (Fourneret and Jeannerod, 1998), suggesting a lack of conscious insight concerning motor action. These data seem to confirm previous findings (Goodale et al., 1986) showing that access to mental representations of action is quite limited from the conscious but not the unconscious perspective.

In sum, converging evidence of non-conscious processing in the brain extending beyond simple visual awareness can be said to exist, in particular for functions linked to cognitive control (see van Gaal et al., 2012; Desender and Van den Bussche, 2012 for review).

1.2 What is metacognition ?

1.2.1 A few definitions.

Parallel to the field of consciousness and almost independently, the question of metacognition has been investigated. What is metacognition? Very broadly, metacognition can be defined as "cognition about cognition", constituting the monitoring, evaluation and control of one's own cerebral processes and behavior. Virtually any process that takes as an input information about another mental process could be described as metacognitive. Metacognition enables us to gain knowledge on our own cognitive processes, allowing us not just to think, but also to know the state of our thinking process. Metacogni-

tion encompasses a slightly distinct idea from the notion of introspection. Introspection which literally means "to look inward" constitutes the examination of one's own thoughts, feelings and conscious state. It is distinguishable from metacognition by its link to consciousness: while metacognition does not necessarily imply the need for conscious experience, introspection assumes the existence of a conscious self who can exert its introspective ability.

1.2.2 The feeling of knowing, the first research on metacognition

Historically, the field of metacognition has been tightly linked to research on memory. The concept of metamemory has been developed to designate our ability to evaluate whether a piece of information can be retrieved from our memory. For instance, metamemory allows us to say with certainty that we know the name of the capital city of France but not of Paraguay. This field of research led to the emergence of the concept of the feeling-of-knowing (FOK). This effect has been shown when subjects are asked to memorize a precise set of items and fail to remember one of them. When the subjects are asked to judge whether, if such item would be displayed, they could recognize it, they are often able to judge quite accurately whether or not they will recognize it (Hart, 1965), even though they still fail to report its identity. A theoretical framework accounting for metamemory and FOK was proposed (Nelson and Narens, 1990) hypothesizing the existence of a "meta-level" feeding from an "object level" which carries information of the first-level about objects stored in memory. Importantly, the "meta-level" is responsible for monitoring and controlling processes occurring during acquisition, retention, and retrieval of memorized objects and plays a role in strategic behaviours such as allocating study time and selecting search strategies.

Beyond the study of memory, several psychologists have investigated the broader question of introspection. In particular, the Sperling experiment on brief visual presentations was particularly striking (Sperling, 1960) in providing compelling evidence for metacognition. In this experiment, subjects were flashed an array of letters for 50 ms (Figure 1.5). On average, subjects were able to report 3-4 letters of the set. Crucially, when an auditory cue just followed the offset of the display and indicated what row should be reported, subjects were then able to report most characters of the row, significantly improving their memory abilities. This counter-intuitive finding was explained by postulating the existence of an iconic memory buffer with a fast temporal decay that can nonetheless be cued in a retrospective manner. It constituted one of the first lines of evidence that the attention of the subject can be oriented to a specific feature of a representation (here a specific location in space), made available after the stimulus has disappeared. This finding was recently extended showing that the display of a cue, either prior to the presentation of the letter array or up to 400 ms after the display, improved performance in recollecting the letter array (Sergent et al., 2013). More importantly for the present thesis, subjective reports were coherent with this result, as noted by Sperling in his original article *"When complex stimuli composed of many alphanumeric characters are displayed with a tachistoscope, subjects enigmatically insist that they saw more than they can remember in retrospect."* (Sperling, 1960). Indeed, Sperling reported that

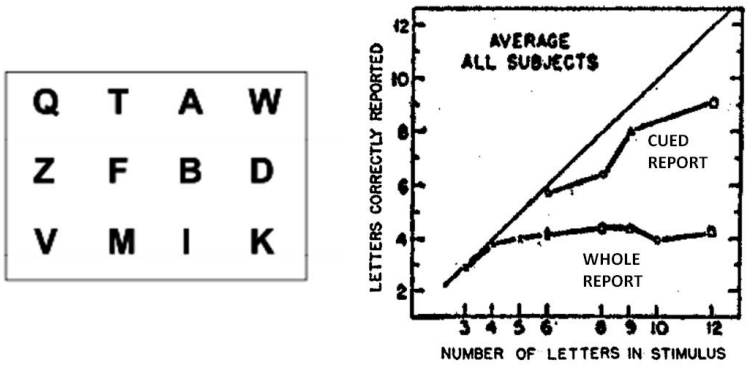


Figure 1.5: The Sperling experiment (adapted from Sperling, 1960). Observers were presented for a brief period of time (50 ms) with a tachistoscopic visual stimulus consisting of either a 3 by 3 or a 3 by 4 array of alphanumeric characters as displayed above. Crucially, when plotting the number of letters reported as a function of the number of letters presented, subject performance was improved when the identity of the letters to report were cued.

while not being able to report all the objects, subjects nonetheless claimed they had seen all of them. Considered in the framework of metacognition and introspection, these reports were in fact true: when properly oriented, any object in the array were indeed available to the subject for report!

Taken as a whole, the results from metamemory studies and the Sperling experiment suggest that subjects can have a good insight into the functioning of their own mental processes, in particular in memory tasks. However, are metacognitive reports always true? In close relation to the feeling of knowing, the feeling of warmth in problem solving corresponds to the impression of being close to solving a problem. In experiments on that effect, subjects were presented with various problems and enigmas. Every ten seconds, subjects estimated with a number between 0 and 10 how "warm" or "cold" they felt about solving the problem (Metcalfe et al., 1986).Surprisingly, it was found that the warmer ratings did not predict that the subjects were close to solving the problem. On the contrary, they were more "warm" reports before an incorrect answer than before correct problem solving. Indeed, what was revealed when looking more closely at cases where problems were solved correctly was that the correct solution emerged suddenly without any prior insight on the part of the subject (Figure 1.6), suggesting that the discovery corresponded to an entirely non-conscious process. In contrast, the gradual feeling of "warmth" observed in incorrect solution may be assimilated to the gradual acceptance of an unsatisfying answer.

1.2.3 Empirical approaches to measure metacognition

Following these initial findings on the accuracy of metacognitive reports, different experimental approaches were proposed. Metacognitive reports can be of several types, concerning almost any cognitive process. A task can be considered as metacognitive when after a decision, an additional report about

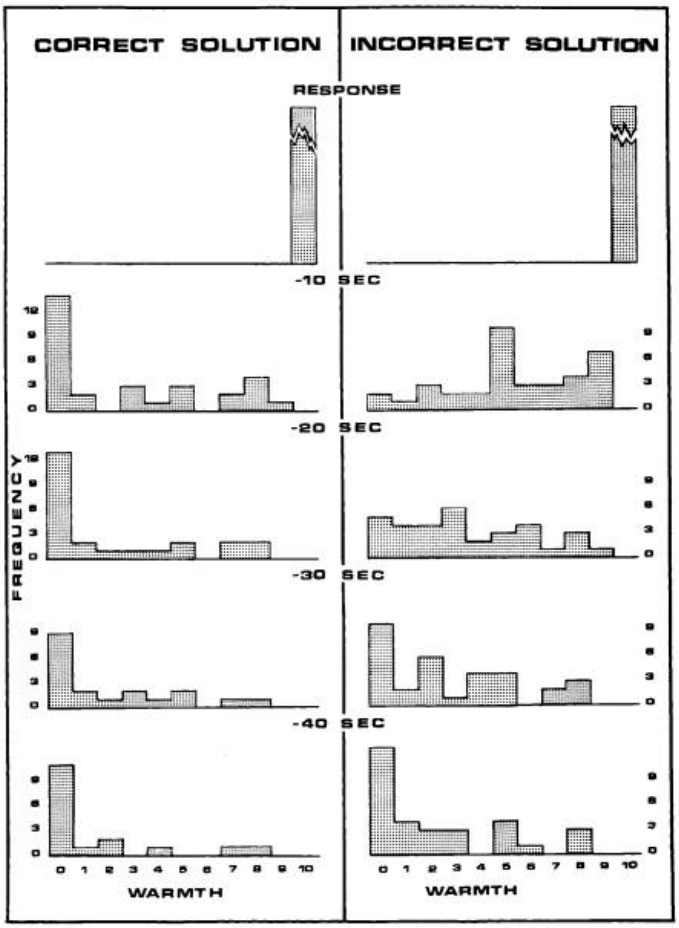


Figure 1.6: Results of feeling of being warm on solving a problem (from Metcalfe et al., 1986). The graphs show that when the solution of a problem was correct, subjects' feeling of being close to the solution increased drastically and suddenly, just before solving the problem (left panel). On the contrary, when the solution found was incorrect, ratings on how close subjects felt they were to the solution increased smoothly, reflecting the gradual acceptance of an unsatisfying solution (right panel)

the initial choice is asked. We talk then of metacognitive or second-order judgment with regards to the first-order response.

Beyond the question of memory, many researchers have investigated the confidence judgments associated with a decision. Historically, Peirce & Jastrow tested themselves on an experiment involving confidence judgments more than one century ago (Charles et al., 1885). Performing a brightness and weight discrimination task, they rated after each of their decisions how confident they were of their choice on a scale from 0 to 3, 0 denoting "absence of any preference for one answer over its opposite" and 3 denoting "as strong a confidence as one could have about such sensations". Using these empirical measures, they found that confidence judgments could dissociate from objective performance in the task. Indeed, even for 0 confidence trials, their accuracy in discriminating the two stimuli was better

than chance. While these very ancient works cannot be regarded as rigorous evidence for the validity of metacognitive judgments, it nonetheless confirms the strong intuition that confidence judgments are valid and carry some distinct information compared to the initial decisions.

The use of a continuous scale for rating confidence is, as we have seen, one of the oldest and simplest. As in Peirce & Jastrow's initial experiment, the method consists of evaluating the certainty of the preceding response on a scale ranging from "Unsure response" to "Sure response". This scale can be continuous, using intermediate levels such as "less sure" or "guess" or consist of a binary judgment. However, as we have seen for conscious perception, many biases can affect the use of such a subjective scale. For example, effects of the overall task difficulty have been documented from an early stage (Gigerenzer et al., 1991) with overconfidence occurring when high confidence judgments are more frequent than actual correct answers or on the contrary under-confidence when the task is actually easier. Note that these types of scales can also be ambiguous. While a "Sure" rating seems to always be associated with a correct response in the subject's mind, "Unsure" ratings may be associated with two different judgments: either the subject has absolutely no clue concerning the accuracy of his or her decision and the judgment is associated with a high level of uncertainty, or on the other hand it corresponds to trials in which the subject thinks he or she made an error, these trials being possibly associated with a high level of certainty concerning the performance. While this issue might be overcome by specifying to the subject which exact interpretation has to be made of this scale, it nonetheless underlines the ambiguity of the confidence task.

Partly to circumvent this issue, other methods linking confidence judgment tasks to betting strategies have been developed. In particular, Persaud and colleagues introduced the post-decision wagering (PDW) method as a way to genuinely measure subject's metacognitive confidence (Persaud et al., 2007). The idea of this method is to ask subjects to bet on the accuracy of their response, larger amounts signaling trials in which subjects are very confident in their response and smaller amounts signaling trials in which confidence is lowest. In its simplest form, the contingency is as follows: if the response is correct, the amount of money wagered is won whereas if it is incorrect, the amount is lost. Persaud and colleagues proposed that not only do such methods allow one to obtain more accurate confidence judgment with the incentives encouraging subjects to provide responses that reflect their internal confidence level, but such measures would also reflect the true conscious experience of the subjects. This argument was supported by the finding that confidence judgment in low visibility task dissociated from objective performance in the task. Applied either to blindsight patient GY or healthy subjects, subjective report of confidence seemed to reflect a special state of subject's perception. In one of the version of the task which required learning some stimulus-response associations, objective performance in the learning task increased rapidly, exceeding chance level while confidence judgment remained constant, reflecting the fact that subjects remained temporarily unaware of their ability to perform the task (Figure 1.7).

This result is particularity surprising considering the specific incentive contingency table used by Persaud et al. (2007). Indeed, the actual optimal strategy would have been to always bet a high wager regardless of performance, as incorrect responses were not penalized. Importantly, in Persaud et al.

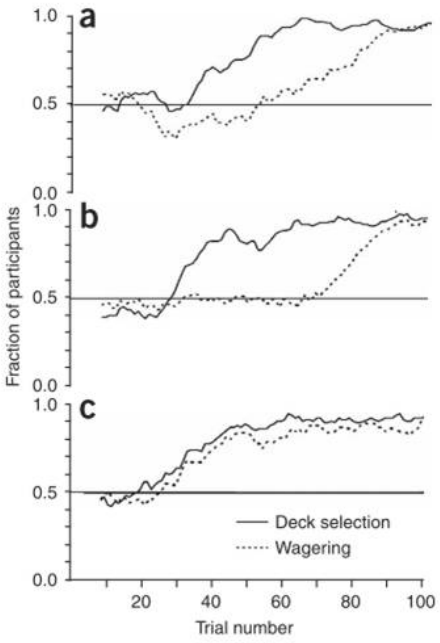


Figure 1.7: Objective performance and wagers in the modified Iowa gambling task (from Persaud et al., 2007). Graphs depict the proportion of subjects that selected the correct pack (solid line) and chose positive wagers (dotted line). Metacognitive knowledge about the task was either not assessed (a), assessed by open-ended questions (b) or assessed by a quantitative questionnaire (c). Results show that accurate wagering started to occur later than the increase of objective performance in deck selection, suggesting a lack of metacognitive knowledge on first-order performance.

initial experiment, subjects did not adopt this optimal but trivial strategy, which would have invalidated their results on awareness. Nonetheless, this possible confound was addressed by other methods, in particular in the field of neuroeconomics. For example, economists proposed an alternative reward system, the Quadratic Scoring Rule (QSR), that uses a contingency table in which the incorrect answers are associated with penalty, making unusable the strategy of continuous high bets (Becker et al., 1964). Even more complex procedures have been used as for example the Lottery Rule (Holt and Smith, 2009) in order to motivate participants to make confidence choices reflecting their true beliefs on their accuracy. However, these paradigms often become very complex, the difficult in understanding the task potentially shadowing the validity of these reports (Hollard et al., 2010).

Cognitive scientists have also used alternative more indirect methods to assess confidence. In particular, one way to determine the uncertainty concerning a response is to give the possibility to not respond for some trials, for instance when the stimuli are judged too difficult. This is particularly useful in animals as it allows one to study neural correlates of confidence without having any verbal subjective report. It has been shown that macaques (Kiani and Shadlen, 2009; Smith et al., 2003; Hampton, 2001), as well as rats (Foote and Crystal, 2007; Kepecs et al., 2008) can be trained to perform such tasks with high accuracy. This technique has also been used on humans (Mamassian and Barthelme,

2009) with subjects being able to judge the difficulty and choose the less uncertain choice among displays of visual stimuli to perform the task. Indeed, subjects in this case were able to measure the visual uncertainty to guide their decision providing an indirect measure of their perceptual confidence.

1.2.4 The neuronal substrates of confidence judgements

Independently of the question of the accuracy of metacognitive judgments, the neural substrate allowing for monitoring our own decisions have been the study of close examination in the last decade. A pioneering study by Wagner (1998) showed that the magnitude of activation of the left prefrontal and temporal cortices during encoding in memory predicts future performance when remembering (Fletcher and Henson, 2001), as predicted by models of PFC as node for second-order judgments. In particular, the neural substrates of metamemory have been investigated in fMRI. Studies of patients with focal lesion in medial prefrontal cortex revealed impairment in metamemory tasks (Modirrousta and Fellows, 2008). Consistent with this suggestion, Chua et al. (2009) found that activity in anterior dorsolateral prefrontal (DLPFC) and lateral prefrontal regions was modulated based on the subjective level of FOK. Overall, metamemory tasks were characterized by greater activity in a large set of regions including medial prefrontal, mid/posterior cingulate, and lateral parietal and temporal regions (Chua et al., 2009).

The neural substrate of metacognitive abilities has also been investigated in decision tasks in non-human primates. In particular, Kiani and Shadlen (2009) investigated the neural markers in decision confidence in rhesus monkeys. Two individuals were trained to make decisions about the direction of moving random dots, trials varying in their level of difficulty. Importantly, the monkeys were rewarded for correct decisions. On some trials, after presenting the stimulus, the monkeys were given the possibility to opt out of the direction decision and go for a "sure bet" for which they received a small but certain reward (Figure 1.8). Interestingly, monkeys used this option more for difficult stimuli than for easy ones, revealing that they were indeed able to assess the degree of certainty to optimize their reward. More strikingly, neurons in lateral intra-parietal cortex, in which evidence is thought to be accumulated for saccadic decisions fired according to the degree of certainty underlying the decision to opt out: intermediate firing of LIP cells were associated with greater likelihood of using the "sure bet" response.

While these results could be interpreted as partially reflecting the property of the stimulus rather than a confidence judgment, other studies have more precisely investigated the neuronal pattern of firing when performing difficult decision task in the period following the decision choice and preceding the reward. In a study in rodents, Kepecs et al. (2008) showed that activity in orbito-frontal cortex (OFC) was predictive of trial outcome, dissociating incorrect from correct decisions, prior to any experimental feedback (Kepecs et al., 2008). In this case, activity reflected either the expected outcome or the second-order judgment of confidence in the response. However, in another study investigating patterns of activity in fronto-polar cortex, (Tsujimoto et al., 2010) found differences between correct and error trials independently of the actual reward, suggesting a coding of the accuracy of the response prior to any feedback. More recently, Middlebrooks and Sommer (2012) found that when monkeys were trained

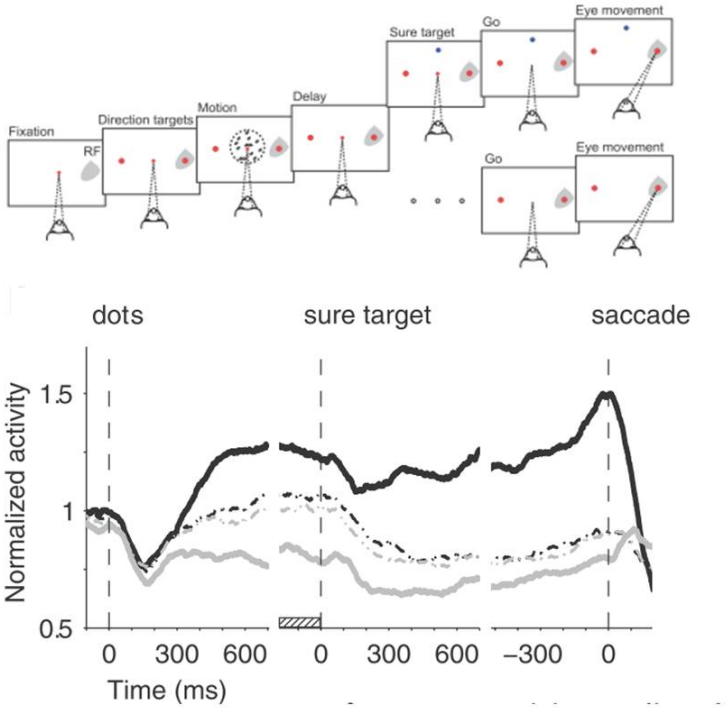


Figure 1.8: Protocol and Results of Kiani and Shadlen, 2009. Top panel shows the task performed by the two monkeys. After an initial fixation time, the motion stimulus appeared for a variable duration. After the target disappeared, the monkey had to make a saccade to one side or the other of the screen to indicate the overall movement direction. For half of the trials (top), an additional option was proposed which consisted of a "sure bet" response. If this option was chosen, monkeys received a smaller reward with 80% probability. Bottom pannel shows the average firing rates of 70 LIP neurons on trials in which the "sure bet" option was presented. The dashed lines show the average neural response on trials in which the "sure bet" was chosen (black and gray, motion towards each direction, respectively).

to bet on their decisions, neuronal activity correlating with bets was found in the frontal eye field (FEF), dorsolateral prefrontal cortex (PFC), and supplementary eye field (SEF).Interestingly, activity that linked decisions to appropriate bets was found exclusively in the SEF.

Several models of decision make use of variables that are tightly linked to the representation of confidence and uncertainty in order to explore patterns of brain activity. In particular, some learning models have been integrated into a Bayesian framework, providing strategies for optimally updating beliefs and using variables tightly linked to metacognitive knowledge. In an interesting study Behrens et al. (2007) tracked and modeled the behavior of participants using a Bayesian learner during a one-armed bandit task in which a choice between two colors had to be made. By changing the ongoing best options at different rate, and in doing so manipulating the uncertainty of the current choice, they showed that the volatility parameter as modeled by the Bayesian learner correlated with activity in the anterior cingulate cortex (ACC). Similar results were found with a maze navigation task in which subjects made

a sequential set of decisions to reach a goal (Yoshida and Ishii, 2006), showing that uncertainty about choice correlated with activity in Brodmann Area 10 (BA10).

While activity related to metacognitive abilities seems to involve several areas, a clear network of regions in prefrontal cortex seems to play a key-role. A particularly convincing study in humans addressed the question of individual differences in metacognitive abilities. Asking subjects to rate confidence in their response after a difficult two-alternative forced-choice task, Fleming et al. (2010) found that the volume of grey matter in the right anterior PFC (Brodmann Area 10), as well as its white matter projection into corpus callosum correlated with the individual ability to rate their performance. Importantly, these results were found by setting the subjects' first-order performance in the task to a common value (Figure 1.9), providing evidence that it reflected the source of the observed differences in metacognitive judgments and not simply the higher-order information contained by the stimulus.

In the same line of research, investigating the causal role of different brain regions in confidence judgments, Rounis et al. (2010) used a paradigm in which they attempted to specifically disrupt metacognitive abilities in healthy subjects. The authors applied transcranial magnetic stimulation (TMS) to dorsolateral PFC (DLPFC) while subjects were making a difficult discrimination judgment on masked stimulus. Subjects performed a two-alternative forced-choice task in which they had to discriminate the relative disposition of two visual stimuli while rating at the same time their subjective visibility ("clear" or "unclear") of the target. After TMS, subjects reported lower visibility levels, even for trials for which they could perform the task correctly, suggesting that they were less able to introspect the accuracy of their decisions. Crucially, signal detection theory analysis confirmed that subjects presented lower metacognitive sensitivity while their first-order discrimination performance remained unimpaired, suggesting a true change in estimating confidence rather than just a modified response bias.

In sum, an important set of studies converge in showing a crucial role of prefrontal regions in metacognitive abilities, in particular DLPFC, ACC and BA10. Importantly this is coherent with findings of these areas as neural correlates of self-reflection (Passingham et al., 2010; Frith and Frith, 2006).

1.3 Models of confidence and error-detection

At the same time as metacognition was studied empirically by investigating the ability to know about one's own mental process and underlying neuronal substrates, the theoretical question of metacognition was also being addressed. Focusing on second-order judgments of confidence in response and the related question of error detection, what kind of theoretical model could account for this type of process? An evident approach to that question is to link first-order and second-order decisions, following the idea that confidence in the response reflects the strength of the underlying evidence used for the initial decision. Indeed, several models have been proposed to model this concept.

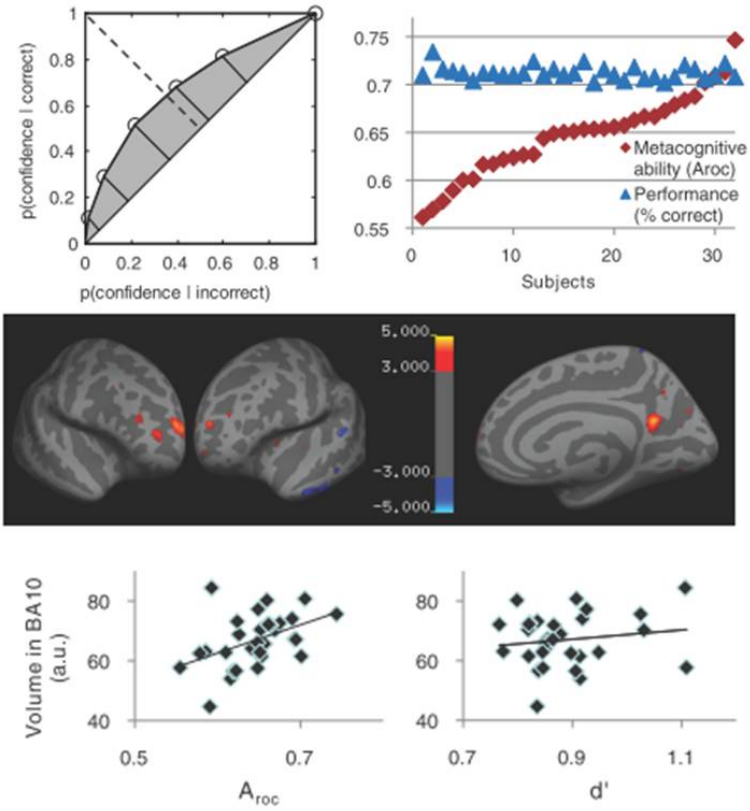


Figure 1.9: Individual differences in metacognitive abilities (from) Fleming et al., 2010). The type II ROC curve from each participant was retrieved and the area-under curve (AUC) was calculated (grey area between the ROC curve and the diagonal, top left graph). Importantly, a staircase procedure ensured that first-order performance (percentage correct, top right graph) stayed constant. However, participants still presented variations in metacognitive accuracy, as seen from the ordered AUC for each subject on top right graph. MRI analysis revealed that grey-matter volume correlated with second-order metacognitive ability (middle panel, T maps for positive correlations with AUC) in right anterior PFC and the left inferior temporal gyrus. Bottom graph shows while grey-matter volume in right BA10 cluster correlated with AUC no such correlation was found for first-order *d'*.

1.3.1 Signal Detection Theory and confidence judgment

We have seen that signal detection theory is a powerful tool with which to assess detection-sensitivity and bias in in the responses to any kind of two-alternative task. This would seem like a good method with which to also assess second-order judgment of error-detection and indeed, various authors have proposed the application of signal-detection theory to second order judgments (Kunimoto et al., 2001; Evans and Azzopardi, 2007). As in classic signal detection theory, we can sort errors and correct trials according to whether they were correctly classified as correct or erroneous, ending up with four categories of trials: "meta-correct" errors, "meta-correct" correct trials, "meta-incorrect" errors and "meta-incorrect" correct trials. This design allows us to draw the same contingency table as for classic SDT (Figure 1.10) as proposed by Kunimoto et al. (2001) and Evans and Azzopardi (2007).

Table 1.
Categorization of responses according to SDT

	Response	
	No	Yes
Blank	Correct rejection	False alarm
Target	Miss	Hit

Table 2.
Kunimoto *et al.*'s proposed categorization of responses incorporating confidence levels, for SDT analysis

Discrimination	Confidence	
	Low	High
Correct	Miss	Hit
Incorrect	Correct rejection	False alarm

Figure 1.10: Contingency table applied on I-order and II-order judgement (from) Evans and Azzopardi, 2007)

Accordingly, it has been proposed to apply *d'* to this table of contingencies, obtaining a value, *a'*, denoting an unbiased measure of error-detection. Importantly, *a'* is computed in a similar way than *d'* following the subsequent equation:

$$a' = Z(h2) - Z(f2) \tag{1.7}$$

where Z is the inverse of the cumulative gaussian distribution, *h2* the II-order hit rate (proportion of correct trials classified as correct) and *f2* the II-order false-alarm rate (proportion of correct trials classified as errors).

However, a detailed analysis of such a measure reveals many difficulties. In particular, such an approach treats type II decisions as a classic decision. However, there is a formal link between type I and type II decisions that cannot be ignored. In a seminal article, Galvin et al. (2003) proposed a complete mathematical analysis of this question. First, let us consider the assumptions of SDT:

1. The evidence about the signal that the observer extracts can be represented in a single number
2. The evidence that is extracted is subject to random variation
3. The choice of response is made by applying a simple decision criterion to the magnitude of the evidence

The first important demonstration made by Galvin is that when projecting the values of II-order decision on the same axis as the initial decision, the distribution of the probability of correct and erroneous

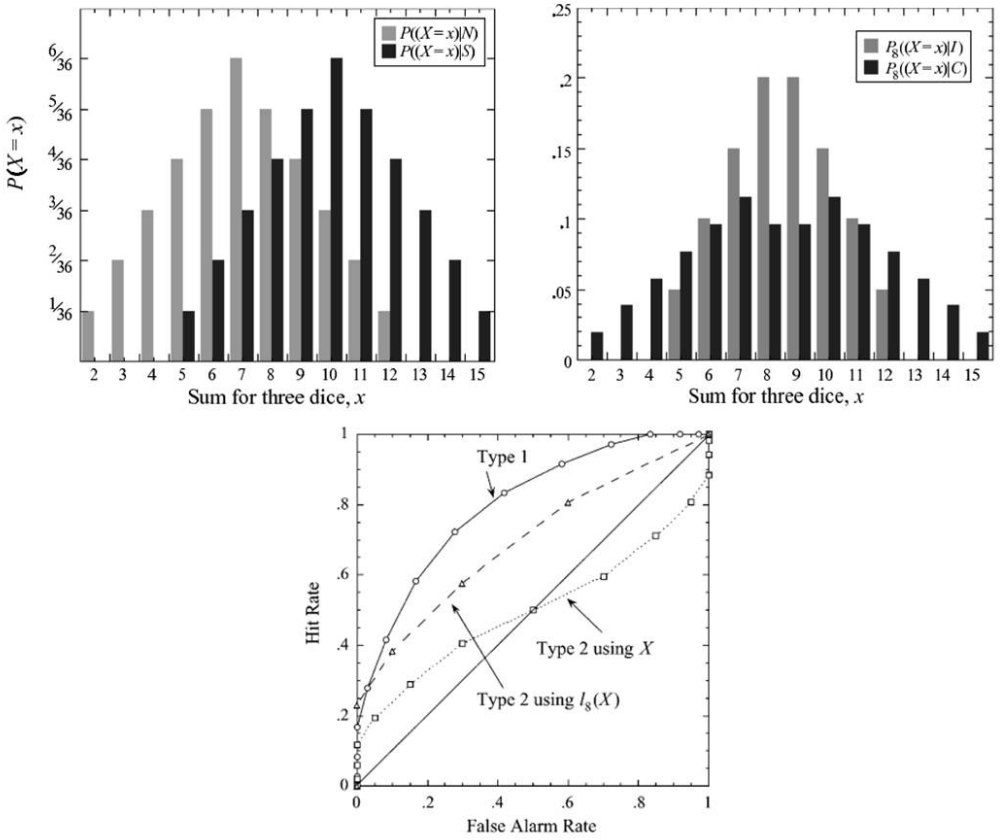


Figure 1.11: Second-order signal detection theory. S. Galvin illustrate their analysis by a dice game: "In the dice game, the experimenter throws three dice (hidden from the observer's gaze) and reports only the sum of the digits on the three upturned faces. Two of the dice are ordinary, but the third has a 0 printed on three sides and a 3 printed on the other three sides. The observer's task on each trial is to use the sum to say whether the strange die has landed with a three facing up (the S event has occurred) or a zero facing up (the N event has occurred)". The distribution on the left represents the respective probability of each conditional event while the distributions on the right represent the distributions of Correct and Incorrect trials for a specified type I criterion (from (Galvin et al., 2003)).)

trial do not follow Gaussian law (Figure 1.11). Indeed, it is intuitive to understand that while distribution for incorrect trials is centered on the *c1* criterion (higher probability of making an error being associated to decision close to the decision threshold) correct trials on the contrary do not follow such a simple distribution. Indeed, while moving further away from the criterion is associated with greater probability of being correct, it remains conditioned by the I-order probability of such a decision value x occurring, with very small and very large values being less frequent.

This result in itself is already a major criticism of directly applying *d'* transformation to II-order as the Gaussian assumption implied by the transformation is violated. Galvin further shows that to simplify the problem of II-order signal-detection, it is necessary to apply a transformation to the I-order decision axis X. In particular, she highlights that the optimal transformation for the decision axis is Type

II likelihood ratio, regardless of the underlying Type 1 probability functions:

$$l(X) = \frac{P(X|S)}{P(X|N)} \tag{1.8}$$

Considering a new decision axis X' and a criterion *c1* for first order decision, what are the additional assumptions needed to model II-order decisions? One intuitive possibility is to simply set an additional criterion *c2* that will be used to classify trials on the correct transformed decision axis X'. Having set the criterion *c2*, it is possible to provide a complete model of II-order decisions and their relation to I-order decisions. From this point, Galvin then demonstrates several key-points amongst which we can consider three major ones:

1. The type I ROC curve provides an upper bound on performance for the type 2 II ROC curve, regardless of the transformation of the decision axes chosen. This statement is quite intuitive: using the exact same piece of information, someone cannot be better in detecting his or her error than he or she was in actually performing the task

2. The crossing point on two distinct type I ROC curves, each of them corresponding to a different type I criteria, will generate the same Type 2 ROC curve when applied to their respective probability functions. This aspect is particularly important from a methodological point of view as it means that when considering the type II distributions, we do not need the exact type I probability function to know the corresponding ROC curve but we just need one that is identical.

3. Type 1 sensitivity (*d'*) and response bias (*c1*) will influence type 2 ROC curve (Figure 1.11, Right panel). This means that even when considering two metacognitively optimal observers, as long as they differ in their type I performance, a difference in type 2 performances may be found, despite the fact that they both have an equivalent detection of their errors.

Overall, this demonstration shows that use of II-order *d'* or *a'* are not appropriate measures for metacognitive performance as they are based on assumptions of the distributions of correct and incorrect trials that do not conform to reality and furthermore are not independent of I-order performance, making the interpretation of these values difficult.

1.3.2 Meta-d'

In the need to develop a method that measures adequately II-order performance, Maniscalco and Lau (2012) developed an alternative measure, *meta-d'*. The goal of establishing such a measure was firstly, to bypass the difficulties due to the specificity of type II distributions and secondly, to obtain a final value of metacognitive performance that truly reflected II-order sensitivity while being independent of I-order performance. To do so, Maniscalco and Lau (2012) proposed the expression of type II

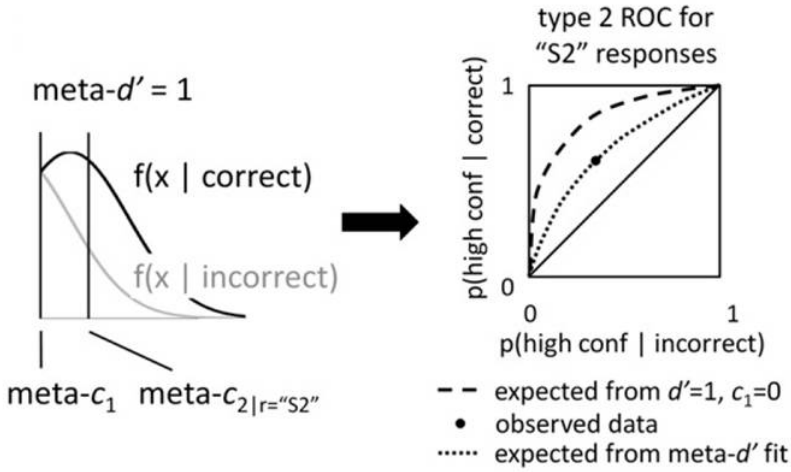


Figure 1.12: Model of meta-d' analysis (from Maniscalco and Lau, 2012). For simplicity, only trials where the subject chose the"S2" response are considered (right portion of the decision). Therefore, the remaining sections of the distribution characterize correct (black line) and incorrect responses (grey line). Fixing *meta-c1* to be equal to c1, it is possible to fit what should have been first-order *d'*, according to second-order distribution.

sensitivity at the level of type I sensitivity. Indeed, as shown by Galvin et al. (2003), it is possible from a given type II distribution to deduce what would have been the associated I-order *d'* value, making the assumption that subjects behave as meta-cognitively optimal observers (Figure 1.12). Using a fitting procedure for II-order criterion, *meta-c2*, and *meta-d'*, assuming that II-order criterion, *meta-c1* is equal to I-order criterion *c1*, the authors performed a maximum likelihood estimation of SDT models allowing the quantification of the likelihood of a given type II data set under a given type I SDT model.

In simpler words, by means of a fitting procedure, this method allows one to find *meta-d'* values which represents what should have been I-order *d'* to produce the empirically observed type 2 data. Therefore, *d'* and *meta-d'* are expressed exactly on the same scale and are directly comparable. If *meta-d'* = *d'*, it means that the subjects simply reported optimally his or her performance according to the initial task performance. On the contrary, if *meta-d'* either exceed or is inferior to *d'*, it indicates that the subject performed the meta-cognitive task with respectively less or more information than the initial I-order task. In this respect, *meta-d'* is intended to measure a relative account of type II sensitivity rather than an absolute one, as it reveals the efficiency of metacognitive judgment. Indeed, it provides a direct measure of the quality of the metacognitive evaluation itself.

This model of the relation between first- and second-order judgments has been shown to be very powerful in explaining how modifications in first-order judgment translate into confidence judgments (Rahnev et al., 2012b; Rahnev et al., 2012a). Moreover, it provides a universal and easy way to assess metacognitive abilities in various experimental conditions (Ko and Lau, 2012; Rounis et al., 2010).

1.3.3 Models of accumulation of evidence for first- and second-order decisions

While this method provides a static view of the relationship between first and second-order decisions, the question of the dynamics of these decisions remains an important point to address. A class of models that provides detailed modeling of dynamics of the decision process is the so-called "evidence accumulation" model (Ratcliff, 1978; Link, 1975) which is derived from random-walks models. In the evidence-accumulation framework, a decision variable (DV) favoring one of the other alternatives is integrated over time. Importantly, the process is subject to random fluctuations, noise being integrated as well as true evidence. Importantly, each piece of evidence is characterized by a drift, favoring one or the other alternative. The decision is made when integrated evidence has reached the threshold for one of the two alternatives. This random walk/diffusion model explains both the final choice and decision times, depending on the model's parameters: the drift rate and the decision threshold. Of course, it also allows one to apply classic signal detection theory analysis to the final decision choice statistics. While in the simplest version of the model, no bias in responding is applied, each of two alternatives being considered as symmetrical, many variations have been proposed in which bias in response is applied. This can be achieved in two main ways: either applying a shift in the starting point of the evidence accumulation process (Link, 1975; Ratcliff and Mckoon, 2009; Diederich, 2006; Bogacz et al., 2006; Voss et al., 2004) or modifying the rate of sensory evidence-accumulation (Diederich, 2006; Ratcliff, 1985). In both cases, more choices will be made in favor of one response than in favor of the other, the reaction-times for the biased response being overall shorter. Using signal detection theory on the modeled responses, it is then possible to obtain a very detailed model of behavioural data.

Interestingly, it is also possible to extend these dynamical models to confidence and error-detection judgments. For example, Pleskac and Busemeyer (2010) proposed a two-stage dynamic signal detection, in which they integrated models of decision making modeling choice and decision time to confidence judgment and second-order signal detection theory. Their model is based on the assumption that in contrast to classic drift-diffusion models, evidence continues to accumulate even after the initial decision. After a fixed amount of time, the level of evidence is tested again and a confidence judgment is made on the basis of the level of evidence at this stage. This simple model provides a good fit with behavioural data of confidence judgment as representing an incrementation of the accumulation process (Figure 1.13).

1.3.4 Dynamical models of error correction

Interestingly, such models also seem to fit the task of error detection. Indeed, the way for an identical system to produce an error and also be able to detect and correct it has long been questioned by psychologists. Rabbitt and colleagues have proposed that indeed error correction might be due to a continuous accumulation of evidence process that might continue to occur even after the initial response has been made (Rabbitt, 2002; Rabbitt, 1966b; Rabbitt, 1966a), providing a model of response competition (Eriksen et al., 1982; Eriksen et al., 1985; Gratton et al., 1988). Indeed, some data obtained in error

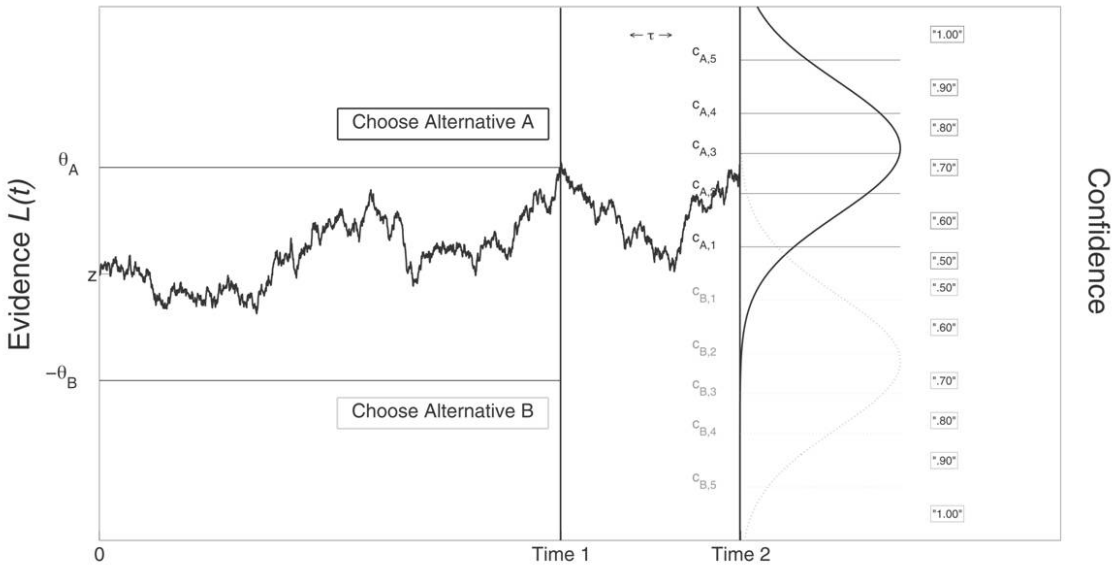


Figure 1.13: A model of accumulation of evidence for first- and second-order decisions (from Pleskac and Busemeyer, 2010). The black jagged line corresponds to the evidence accumulation process for choosing either response A or B. After the decision threshold for response A is crossed at Time 1, evidence continues to accumulate. The confidence judgment is produced after a fixed time interval determining the level of evidence at that Time 2. Gaussian vertical distribution reflects the distribution of evidence at Time 2, when the subject has correctly chosen response A. Confidence level is based on the division of this distribution by confidence criteria.

correction paradigms (Rabbitt and Vyas, 1981) suggested that error correction rate was related to the amount of processing time allowed by the stimulus presentation. Therefore, a model of error correction based on continued processing after the initial erroneous response seems to be a plausible explanation of its mechanism.

Such a model has been updated in a recent article from Resulaj et al. (2009). Using a robotic interface, subjects had to respond to a stimulus comprised of moving dots with variable motion strength. Interestingly, movement towards the correct answer were not always direct: their hand moved initially in the direction of the incorrect response before being corrected suggesting a late "change of mind" concerning the decision. The authors proposed a model (see Figure 1.14) that explains how evidence is accumulated in such simple decision making task that nonetheless account for the making of errors and their correction.

In particular, the authors proposed that subjects do not use the totality of the available information to make their initial decision but process the rest of the information in a later stage to correct or maintain their decision. The precise mechanism proposed is comparable to the classic drift diffusion model in which the decision variable accumulated up to a specific threshold is crossed, simulating the initial reaction time. Importantly, further accumulation occurs on the evidence still in the processing pipeline. If the accumulated evidence reaches an additional threshold, "the change-of-mind bound" corresponding

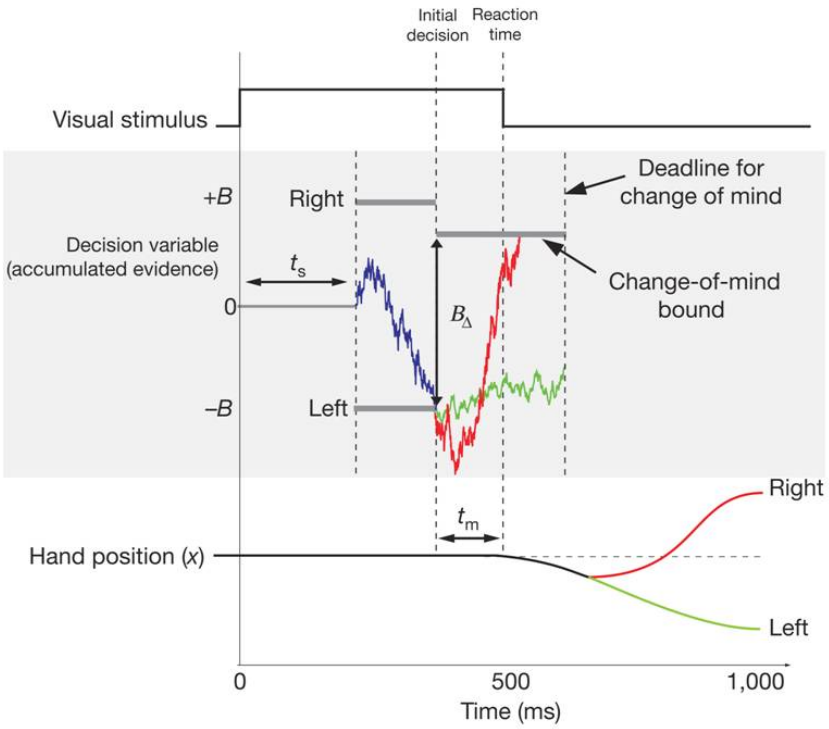


Figure 1.14: Model of Change of Mind (from Resulaj et al., 2009). The blue line represents the evidence accumulation process used to decide between a Left (bottom) and a Right (Top) response. Initially, the "Left" threshold is crossed and therefore a left movement is initiated. However, accumulation continues to takes place and can result either in the confirmation of the movement (green line) or in the correction of the decision (red line) if the decision variable crosses an additional change-of-mind bound.

to the opposite decision then the decision is reversed and the motor action is corrected. While this model seems to provide a powerful explanation of the mechanisms leading to the making of an error and its subsequent correction, it is questionable whether such a mechanism can account for the rapidity of some error correction processes. For example, previous work by Rabbitt (2002) showed that subject could correct their errors very accurately in less than 150 ms after the first erroneous response in a sort of delayed correct response. Due to the very fast occurrence of such a correcting motor response, the question remains as to whether a serial process such as the one described here might be sufficient to account for all processes of error correction. Alternatively, it has been proposed that evidence accumulation process for the correcting response might run in parallel with the normal response process (Rabbitt, 2002), explaining the automatic aspect of error correction that is sometimes difficult to inhibit.

We have seen that several theoretical models have been provided to explain how initial decision and second-order judgments of confidence can be related, both on a static and on a dynamic point of view. In the next section, we investigate the specific metacognitive task of error detection and discuss how it might relate to these models.

1.3.5 Alternative models of confidence judgments

Radically distinct models have been proposed for confidence judgments. In a recent study, (Zylberberg et al., 2012) investigated how confidence ratings from a continuous scale were influenced by first-order evidence. Subjects performed two experiments: a motion-discrimination task on random-dots and a luminance discrimination task on pairs of pseudo-gabor patches. Interestingly, their experiments revealed two important empirical findings: firstly, that confidence judgments appear to be correlated with the first moments of accumulation of evidence rather than with later stages of decision process and secondly, that evidence of the non-selected choice do not appear to be taken into account when determining confidence, as if confidence reflects only the "positive evidence" accumulation process. The authors discussed their findings in the framework of the different theoretical models of decision. Since results showed that only evidence about the chosen stimulus was used to produce confidence judgments, they seem difficult to reconcile with random-walk models in which the decision choice is made based on the "difference" between signals favoring one or the opposite response as proposed in the models described above (Pleskac and Busemeyer, 2010; Resulaj et al., 2009). Rather, their findings speak in favor of "race" models in which evidence about each of the two alternative responses are accumulated separately. Furthermore, the authors suggest that a model of confidence relying uniquely on decision-time to determine confidence could account for their data. Indeed, decision-time, referring to the time taken to reach the decision threshold, reflects for each trial the slope of evidence accumulation, providing a measure of how "easy" the decision was to make.

At the same time, Yeung and Summerfield (2012) proposed an alternative model for confidence judgments that takes into account the reliability of evidence. They proposed that instead of only considering the mean of the strength of the decision variable, confidence judgments also evaluate its variance. According to this view, the decision variable would be the probability distribution of the evidence accumulation process that evolves across time (Figure 1.15). In this framework, the variance of the distribution reflecting the noise in the accumulation process itself would be entered as a factor in the final confidence judgment, providing a representation of evidence reliability. Importantly, such values would also be available in a continuous manner, at any point in time (Yeung and Summerfield, 2012).

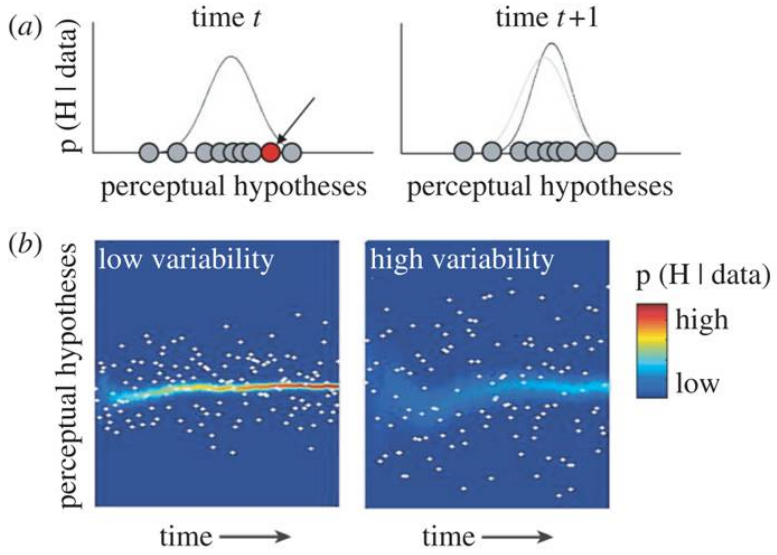


Figure 1.15: Model of confidence judgement based on both the mean and variance of evidence accumulation(from Yeung and Summerfield, 2012). In the top left panel, each grey dot represents the level of accumulated evidence at different time-points. The red dot corresponds to the new level of evidence at time *t*. The grey line shows the overall posterior probability distribution $p(H \mid data)$ after a given time *t*. At time $t + 1$, a new distribution taking into account the new sample data point can be retrieved. Confidence corresponds to the precision of the distribution," i.e. the reciprocal of its standard deviation". In bottom graph, we can see how the confidence value $p(H \mid data)$ evolves through time (x-axis), as new sample updates the posterior probability distribution. On the left graph, low variance in the evidence accumulation process corresponds to a rapid increase in the precision of the posterior distribution for hypothesis H while for high variance (right graph), precision increases more slowly and remains overall lower.

CHAPTER 2

Error-detection, a simple metacognitive task

"Oops!" Who has not had the experience of making an error? Detecting our own error is probably the most intuitive metacognitive judgment that one can make and it has been widely studied by many cognitive scientists. In particular, the question of whether error-detection can operate non-consciously has been investigated in many studies. In this section, we present the neural substrate of error detection and its relation to consciousness as well as how it is altered in some pathologies, especially schizophrenia.

2.1 A brief review on error detection

The question of performance monitoring has been an important subject of research for several decades. Why do we make errors? How do we detect them? How do we correct them? What are the consequences of an error on future behavior? Interestingly, the subject of error was first investigated from the point of view of post-error adjustment rather than the causal mechanisms leading to the making of the error. In particular, several authors studied the mechanisms of error detection and error correction (van Veen and Carter, 2006; Yeung et al., 2004; Danielmeier and Ullsperger, 2011), as well as post-error adjustments, in connection with the more global topic of cognitive control.

Pioneering research on this subject was led by Patrick Rabbitt in the mid 60s (Rabbitt, 1966b; Rabbitt, 1966a). In particular, this first body of research focused on reaction times before, during and after making an error. In this first work, Rabbitt showed that errors and error corrections were characterized by faster reaction times (RT) compared to correct trials. In particular, he showed that error correction could occur in a very fast and automatic manner, a few hundred milliseconds after the first incorrect response. Furthermore, his work showed, for the first time, that trials following errors were characterized by a slower response time. This work played a central role in further research as it highlighted the special processing of motor errors by the brain. Several studies since then have confirmed this finding (Laming, 1968; Notebaert et al., 2009; Núñez Castellar et al., 2010; Danielmeier and Ullsperger, 2011; Strozyk and Jentzsch, 2012) that in some conditions, trials following errors present much slower RT, a phenomenon called Post-Error Slowing (PES). The overall pattern shows that errors are associated with faster response-times than correct trials but are followed by an immediate slowing of RT, progressively decreasing while the error becomes more distant in time. The same pattern appears for response accuracy, which rises after errors as can be seen in Figure 2.1.

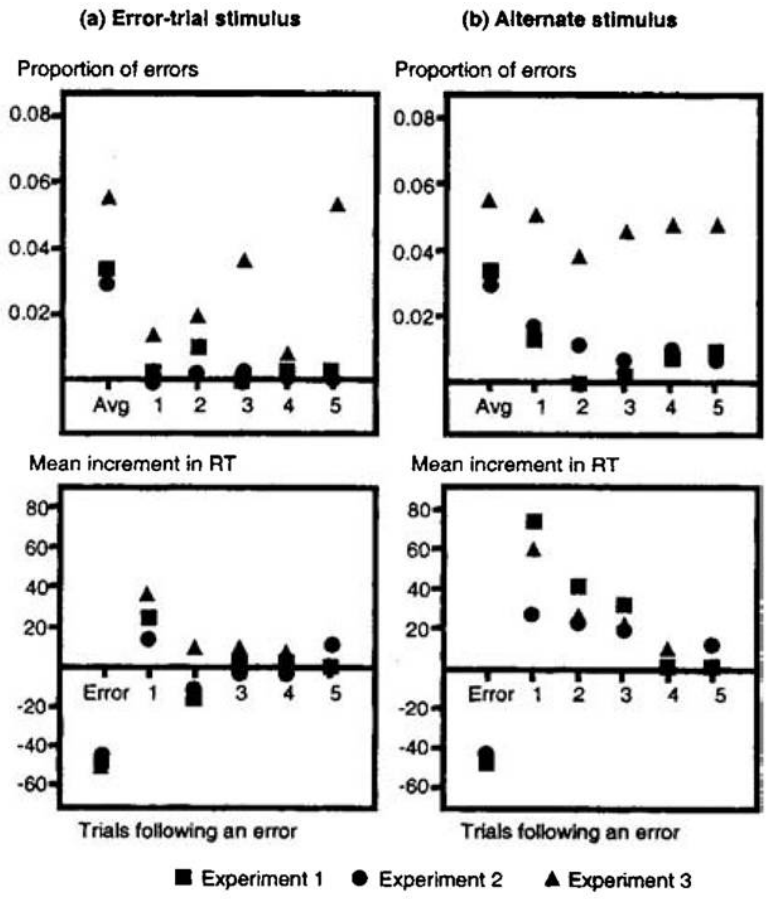


Figure 2.1: Results of post-error slowing and proportion of errors after an error (from Laming, 1968). For each trial following an error (x-axis) the accuracy (top) and the reaction-times (bottom) are plotted, in cases where the stimulus from the error trial was repeated (left graphs) and trials where the alternative stimulus occurred (right graphs). We can see from bottom graphs that error corresponds to the fastest reaction-times and are followed by an immediate slowing down of the RTs, which progressively decrease again when the error trial becomes more distant in time.

This phenomenon has been discussed in terms of cognitive control and several models have been proposed to explain it (Laming, 1968; Laming, 1979b; Laming, 1979a). In particular, PES has been linked to top-down control and maintenance of accuracy (Botvinick et al., 2001), as predicted by models of conflict monitoring. According to this view, errors which are associated with greater conflict between the executed and the required response lead to a reduction in response priming. Such an effect results in slower and more accurate responses for a short period, before once again reaching a period of low conflict in which accuracy decreases while RTs become shorter. Importantly, conflict can also occur in correct trials without systematically leading to an erroneous response. Therefore, fluctuation of RT can also be recorded for correct trials, depending on the amount of conflict in each trial.

In a similar vein, it has been proposed that PES could be associated with the remaining motor inhibi-

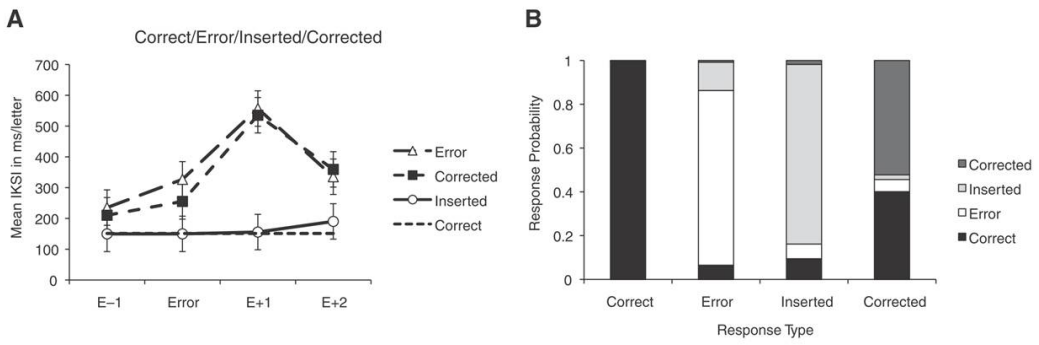


Figure 2.2: Artificially corrected errors nonetheless trigger post-error slowing (from Logan and Crump, 2010). Left graphs show the mean response time (inter-key stroke interval in milliseconds per letter) plotted from the trial preceding the error (E - 1), to the second trial after the error (E+1, E+2) for the four types of trials (correct responses = correct, actual errors = error, inserted errors = inserted and corrected errors = corrected). After the error, even when the error is corrected, we can observe a slowing-down of the responses. However, such an effect does not exist for inserted errors. Right graphs depict the perceived performance (correct, error, inserted error, or corrected error) for the four types of trials. Importantly, the results reveal that subjects do not detect all the corrected and inserted errors.

tion of the incorrect response (Marco-Pallares et al., 2008; Ridderinkhof, 2002). Indeed, the amount of PES correlates with activation of a known network of response inhibition (Marco-Pallares et al., 2008; Kühn et al., 2004). Interestingly however, post-error slowing has also been associated with related task settings such as error frequency and error awareness. In particular, some authors investigated the link between post-error slowing and the overall error-rate (Barceló et al., 2006; Notebaert et al., 2009). Rather than being linked to cognitive control mechanisms, PES may simply reflect the fact that errors are much less frequent than correct trials, the slowing-down being related to the detection of a rare event (Barceló et al., 2006; Notebaert et al., 2009). In this framework, Notebaert et al. (2009) used a color discrimination task paradigm in which they manipulated the frequency of errors by adjusting on a trial-by-trial basis the brightness of an image on which subjects performed the task. Crucially, post-error slowing was observed when errors were infrequent. However, when they became more frequent, correct trials corresponding to rare events, slowing was observed after correct trials and not after errors. These results were confirmed by other studies showing an increase in post-error slowing when errors were less numerous and speed was emphasized over accuracy (Ulrich and Szymanowski, 2004). Importantly, these results seem difficult to reconcile with the conflict monitoring view, without assuming additional brain process related to the tracking of the ongoing task. They might however be better understood when confronted with other findings regarding error awareness. In particular, several authors reported that post-error slowing only follows errors that are detected (Nieuwenhuis et al., 2001; Wessel et al., 2011) or that it is strongly reduced in undetected errors (Cohen et al., 2009). It is therefore possible that when errors are very frequent, subjects are mostly unaware of them, failing to trigger mechanisms of conscious error-detection.

However, a very striking study recently contradicted this view (Logan and Crump, 2010). Investigating the performance of skilled typists, the authors used a simple word computer writing paradigm.

Importantly, they manipulated the screen output of the words typed, inserting or on the contrary correcting errors made by the subjects. Their results show that when asked to report their errors, typists blame themselves for errors that were artificiality inserted and took credit for the corrected errors. In other words, subjects systematically claimed responsibility for the words as they appeared on the screen, revealing a strong illusion of authorship even when their behavior did not match the result. However, different results were found for their typing rate, revealing no effect of these illusions. Indeed, subjects presented post-error slowing after errors that appeared corrected but not after inserted errors. The authors suggested that these findings provide evidence for the existence of two error-detection processes sensitive respectively to the output of the action (here the appearance of the words on the screen) and the actual action. According to their findings, post-error slowing would be sensitive to the action itself, independently of the awareness of the action. While more work will be needed to understand how to reconcile all these findings, they nonetheless demonstrate the potential of such a measure as an index of action monitoring.

In addition to post-error slowing, various other behavioural adjustments have been observed following the making of an error. In particular, several studies have investigated how errors play a role in learning and improvement of performance. In particular, post-error improvement of accuracy has been described in several studies (Marco-Pallares et al., 2008; Maier et al., 2011; Danielmeier et al., 2011). While this finding appears to be less reproducible according to the task (Hajcak and Simons, 2008), the impact of error-related brain activity and post-error behavior on the learning process has been highlighted by many studies (Klein et al., 2007b). However as this aspect is not directly related to the subject of the present research, we will not discuss these findings in any more detail here.

2.2 The Error-Related Negativity: a cerebral marker of error detection

Cognitive scientists have investigated the neural correlates of error making and neuroimaging data has widely contributed to the understanding of error processing and cognitive control. In particular several research teams reported a marker of neural activity specific to errors in the early 90s (Dehaene et al., 1994; Gehring et al., 1993; Falkenstein et al., 1991). EEG studies revealed that when performing a task, erroneous motor responses are followed by a specific negative ERP component occurring between 50 and 150 ms (see Figure 2.3) after the wrong key-press. This error specific ERP named Error-Related Negativity (ERN ou Ne) has a characteristic fronto-central distribution (see Figure 2.3), peaking maximally at electrode FCz. Importantly, it is followed by a positive component (Pe) occurring between 150 and 250 ms after the motor response and which lasts for several hundred milliseconds, with a similar but slightly posterior topography.

Importantly, the ERN has been observed in various experimental conditions, independently of task settings, stimulus modality (Falkenstein et al., 2000)and motor response (Holroyd et al., 1998). Interestingly, it has been shown that an ERP with the same topography as the ERN may be elicited simply by the observation of someone else making an error (Schie et al., 2004). Furthermore, a very similar

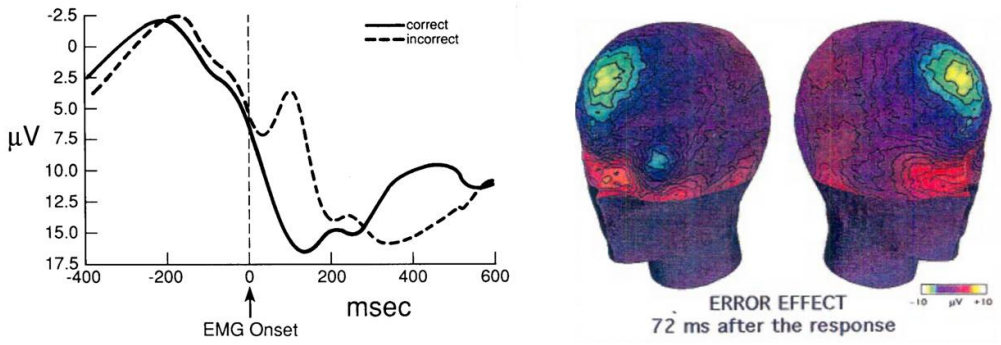


Figure 2.3: The Error-related Negativity (from Gehring et al., 1993) and its topography (from Dehaene et al., 1994)

signal called the Feedback-related Negativity (FRN) is observed when feedback is provided on motor performance. Finally, several studies have revealed that an ERN of very small amplitude is present even after correct trials (Luu et al., 2000; Vidal et al., 2000). The discovery of this component designated as the Correct-Response Negativity (CRN) suggests that the ERN is not completely absent when no error is committed and therefore reflects a process occurring in both correct and error trials. Overall these results provide converging evidence that the ERN might be part of a more generic system related to performance monitoring.

2.2.1 Factors influencing the ERN amplitude

The ERN seems to be observable in various situations, irrespective of the task or the sensory modality. Falkenstein et al. (2000) showed, for example, that an identical ERN was evoked when performing a task on visual and auditory stimuli. Similarly, the ERN was shown to be present regardless of the motor action performed, being triggered by hand as well as foot actions (Gehring and Fencsik, 2001; Holroyd et al., 1998). The ERN has been observed in a great variety of task sets (Falkenstein et al., 2000), such as the eriksen flanker task (Falkenstein et al., 1991; Gehring et al., 1993), stroop (Riesel et al., 2013; West and Travers,), go no-go tasks (Riesel et al., 2013; Bates et al., 2002) and the number-comparison task (Dehaene et al., 1994). Importantly, the ERN is present in the absence of explicit feedback.

Interestingly, it was initially found that the ERN is reduced when time-pressure increases (Falkenstein et al., 1991; Falkenstein et al., 2000; Gehring et al., 1993). In particular, the ERN appeared to be larger when accuracy was emphasized (Gehring et al., 1993). However, further work has shown that time-pressure in itself does not influence the amplitude of the ERN (Falkenstein et al., 2000) Rather, it was a confounding factor- the overall error-rate- that modified the amplitude of the ERN. In particular, Falkenstein et al. (2000) showed that when errors were averaged separately according to RT, no amplitude difference were found. On the contrary, when subjects were split according to their error-rate while verifying that overall RTs remained identical, the ERN was reduced in the group

that made more errors. Other studies however did not replicate these findings (Pailing et al., 2002; Pailing and Segalowitz, 2004a) showing that the important factor in ERN amplitude variations is the subjective rather than the objective difficulty of the task (Pailing and Segalowitz, 2004a; Scheffers and Coles, 2000).

Interestingly, the ERN amplitude has also been shown to correlate with post-error adjustments. In their original paper, Gehring et al. (1993) found that a greater ERN was associated with a higher probability of correction, as well as slower RTs on the following trial, suggesting that the ERN might correlate with the control of the following responses. Similarly, some authors showed that when splitting trials according to the speed of error-correction, errors that were corrected in a fast-manner were associated with greater ERN amplitude (Rodríguez-fornells et al., 2002). Moreover, the timing of the ERN has also been linked to the timing of error correction (Fiehler et al., 2005). Indeed, uncorrected errors were associated with a delayed ERN, the peak occurring 15 to 20 ms later. Late ERN peak has also been related to decreased attention and a lower correction rate (Falkenstein et al., 2000). This was interpreted as the result of an impaired response determination process, suggesting that the ERN might not be time-locked to the motor response itself but rather to the computation of the correct response.

2.2.2 Location of the origin of the ERN

The topography of the ERN consists of a fronto-central distribution. Using a dipole fitting procedure with one single dipole, the origin of the ERN was first located by (Dehaene et al., 1994) in the Anterior Cingulate Cortex (ACC). This finding was further replicated using similar dipole models (Holroyd et al., 1998; Gehring et al., 2000; Alain, 2002; Munro et al., 2007; O'Connell et al., 2007; Van Veen and Carter, 2002; Vlamings, 2008; Vocat et al., 2008). A study using intra-cerebral ERP recordings from epileptic patients tended to confirm this finding showing sites responding specifically to errors in the ACC (Brazdil et al., 2002), in its more rostral part. However the study also found many local generators of the ERN in mesio-temporal and dorsolateral prefrontal cortex, raising doubts regarding the specificity of the observed effects. The authors suggested that the activation in these regions might be linked to other error processes such as emotional value or post-response adjustments, highlighting the multiple brain regions responding to the making of an error.

A powerful study by Debener et al. (2005) nonetheless confirmed the involvement of the ACC in the ERN, using both fMRI and EEG techniques simultaneously (see Figure 2.4). In addition to dipole fitting the source of the ERN, which was found again in ACC, they used simultaneous EEG-fMRI recordings to study single trial amplitude of the ERN in EEG data and the BOLD signal in response to errors. Using an EEG informed analysis of fMRI data, the authors showed that the activity in the Rostral Cingulate Zone (RCZ corresponding to the rostral area of the ACC) increased with the amplitude of the ERN (Figure 2.4).

Many fMRI studies have shown activity linked to error monitoring in the ACC (Botvinick et al., 2004; Veen and Carter, 2002; Brown and Braver, 2005; Cohen, 2010; Chevrier and Schachar, 2010),

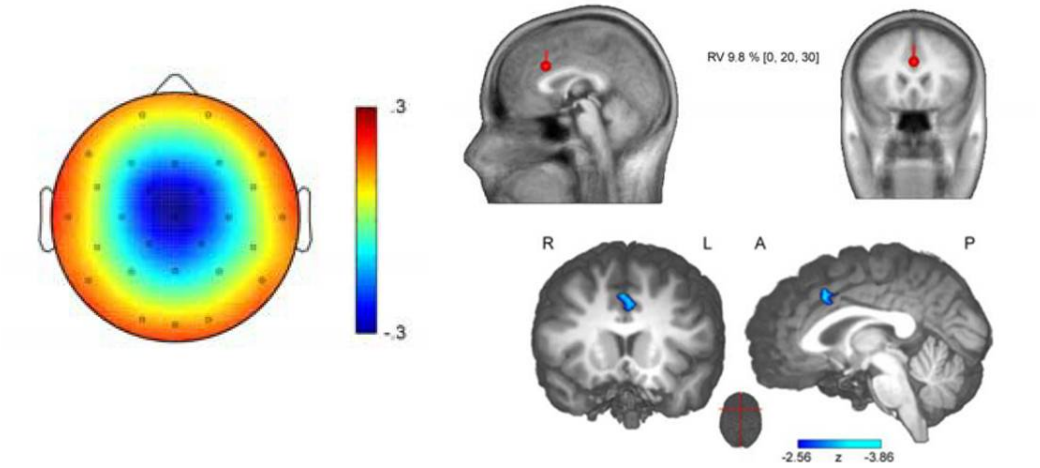


Figure 2.4: Location of the generator of the ERN (from Debener et al., 2005). Left panel shows the topography of the ERN. On the top right panel is the source reconstruction of the dipole of the EEG signal and on the bottom right the EEG informed fMRI reconstruction of the origin of the ERN

making it a very plausible candidate for generating the ERN. However, several studies have suggested alternative explanations. In particular, motor regions have been shown to also participate in the generation of the ERN (Ullsperger et al., 2003; Herrmann et al., 2004) and it has been suggested that the ERN may be generated in the Brodman area n°6 and in particular in the Premotor or Supplemental motor area (SMA) as well as in caudal region of the ACC (Dehaene et al., 1994). When investigating the magnetic equivalent of the ERN in magneto-encephalographic recordings (MEG), Miltner et al. (2003) found that it was generated by the ACC. However, the authors found a great variability of this component across subjects and the data seemed less clear than that obtained with EEG.

More recent studies using methods of distributed source reconstruction found a slightly more posterior origin for the ERN(Herrmann et al., 2004; Aarts and Pourtois, 2010; Hochman et al., 2009). A recent study by Agam et al. (2011) confirmed this finding. Combining simultaneous MEG-EEG, as well as fMRI, with state of the art methods of forward and inverse modeling, the authors found that the posterior region of the cingulate cortex was primarily generating the ERN but could have been missed in fMRI studies due to its different spatial sensitivity compared to electro-physiological measures. Such a finding is particularly interesting as it shows the potential differences that exist between measures and therefore the advantage in combining MEG and EEG for source reconstruction. While the debate on the true generators of the ERN is still lively, a clear network involving motor regions, in particular pre-SMA, posterior and anterior cingulate cortex, and possibly precuneus is thought to participate in generating the ERN.

Several studies tried to more precisely investigate the difference in the origin of the Ne and the Pe. The majority of the results seemed to be in favour of a slightly different generator for the Ne and the Pe. One study by Brazdil et al. (2002) with intracranial recordings indeed suggested that the ERN and the

Pe have the same origin. However more recent work supports the idea of slightly different sources of the ERN and the Pe, in more posterior regions than the ACC (Vocat et al., 2008; Veen and Carter, 2002; O'Connell et al., 2007). These differences could be potentially explained by the fact that the Pe, as the P3 component can be decomposed into several components: an early component occurring 150 to 300 ms after the erroneous motor response which has the same origin as the ERN and a later component (300-600 ms) originating in a more anterior region of the ACC (Endrass et al., 2007; van Veen and Carter, 2006).

2.2.3 Functional Role of the ERN

Following the discovery of the ERN, several studies have tried to determine the exact cognitive processes it reflects. While the ERN was initially thought to reflect the detection by the brain that an error had occurred (Falkenstein et al., 1991), the discovery that it was not completely absent in correct trials forced this initial framework to be revisited (Falkenstein et al., 2000). Currently, three main theories have been developed regarding the significance of the ERN: the "mismatch" or the comparison model, the conflict monitoring model and the reinforcement-learning model.

In the comparison model or mismatch theory, the ERN reflects the comparison process between the actual and the required response (Coles et al., 2001; Falkenstein et al., 2000; Gehring et al., 1993; Scheffers and Coles, 2000; Scheffers et al., 1996). According to this view, the stimulus to which the subject responds continues to be processed after the response, even in the case of a correct response. The representation of the correct response associated with the stimulus is computed, sometimes even after the initial response, and is compared to the efferent copy of the motor response. In this framework any mismatch or discrepancy between required and executed actions would trigger an ERN. The existence of a small negativity in the correct condition would then reflect the evaluation of the correctness of the motor response, after the additional processing of the stimulus. In contrast, errors would result from responses that did not make use of complete processing of the stimulus, producing fast erroneous responses. Some authors proposed a variant of this model in which the onset of the ERN depends on the moment the correct response is computed (Falkenstein et al., 2000). While not many studies have directly investigated the trial-by-trial variability concerning the onset of the ERN, the various timings that have been reported in the literature suggest that it is quite consistently locked to the motor response onset, therefore making it difficult to conclude on the validity of this hypothesis. Importantly, the comparison model predicts that the ERN should vary in a trial-by-trial manner according to the mismatch between the actual and required response and reflecting the amount of evidence available concerning both types of information. While this aspect has not been directly investigated by many studies, evidence that the ERN varies with the level of uncertainty concerning the correct response has indeed been found (Scheffers and Coles, 2000; Pailing and Segalowitz, 2004a; Hughes and Yeung, 2011).

The second theory that tried to account for the existence of the ERN is the theory of conflict detection. According to this hypothesis the ERN reflects the conflict between two con-

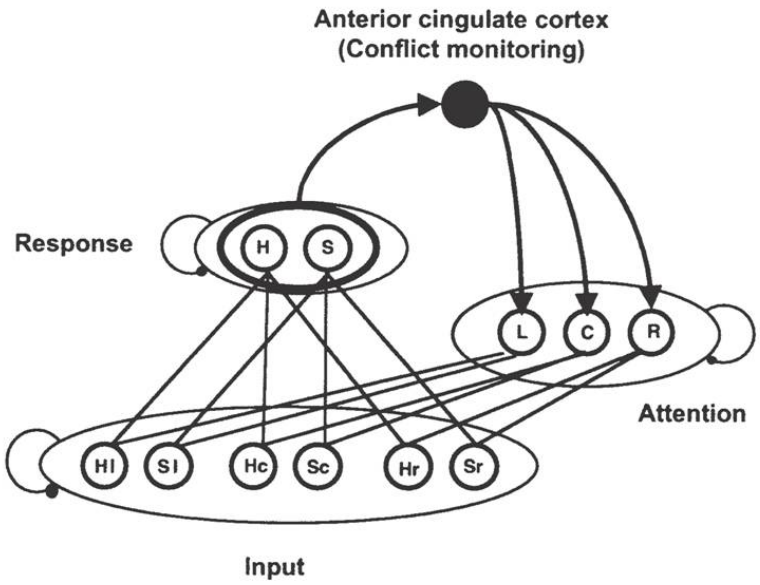


Figure 2.5: Model of conflict monitoring (from Yeung et al., 2004). See text.

tradictory streams of information reflecting respectively the erroneous fast response and the slower correct response. Importantly, this theory is associated with precise computational model of conflict evaluation which has been linked to numerous research in fMRI (Veen et al., 2001). Indeed, ACC has been shown to activate in many studies independently of performance when multiple responses are possible and are in competition (Veen et al., 2001; Veen and Carter, 2002; Van Veen and Carter, 2002; Botvinick et al., 2001; Botvinick et al., 2004; Kiehl et al., 2000). This effect is particularly salient when using paradigms such as the Eriksen flanker task. In this task, the target stimulus which indicates the required motor response (for example an arrow > pointing to the side of the response hand) is embedded in an array of distractor stimuli that are associated with the opposite motor response (for example <<><<). This paradigm allows one to distinguish congruent trials in which the target and the distractors evoke the same motor response, from incongruent trials in which they both evoke opposite responses. In this type of paradigm, ACC is seen to activate preferentially in incongruent trials in which conflict is high (Botvinick et al., 2001; Carter et al., 1998). According to the conflict monitoring model, detection of conflict will then lead to the triggering of further regions associated with cognitive control in order to shift attention resources and increase top-down control (Botvinick et al., 2001; Botvinick et al., 2004). Different connectionist models have been proposed to account for the conflict monitoring hypothesis and the ERN. In a very thorough article, Yeung et al. (2004) proposed a version of the model that integrates the different results concerning the ERN (Figure 2.5). This model simulates the response in a classic Eriksen flanker task where the target stimulus is a letter (H or S)

appearing on the center of the screen and surrounded by an array of flankers, congruent or incongruent with the target letter. Inputs consist of the four possible stimuli represented as patterns of activity across different input units of the model and subject to noise. The model is composed of three layers: an input layer consisting of six letter units, a response layer with one unit for each responses, and an attention layer with units corresponding to each location in the letter array (Yeung et al., 2004). Importantly, the weights between layers are bidirectional, the inhibitory weights between the unit of the same layer corresponding to the competition between responses. In this framework, conflict is computed as the product of activity of the two response units weighted by the connection strength between the two, as generalized from the following conflict equation:

$$Conflict = -\sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j w_{ij} \tag{2.1}$$

where a denotes the activity of a unit, w the weight of the connection between a pair of units, and the subscripts i and j are indexed over the units of interest (Yeung et al., 2004). Intuitively, we understand from this model that when only one response unit is active and the other inhibited (with an activity close to 0), the conflict is low while when the two units are active, the conflict is maximal. In this framework, RT are modeled as the number of cycles necessary to reach a given threshold, plus a time-constant that might correspond to perceptual time not linked to the decision. What are the dynamics of conflict according to this model? Yeung et al. (2004) showed that because of this additional noise in the input units, the incorrect response is sometimes triggered before the stimulus is fully processed, modeling the occurrence of an error resulting from a fast guess. In this case, response conflict reaches its maximum in the period following the response, where the activation of the correct response emerges and conflicts with the remaining activity in the opposite response unit. Interestingly, such a pattern is very transient, as the high conflict situation associated with the co-activation of the two response units is incompatible with inhibition between the two units. Therefore, the unit with the maximum activity (in general the one corresponding to the correct response) quickly inhibits the other and dominates the response units pattern of activation. However, in correct trials a different dynamic is observed and conflict is observed before the onset of the motor response. In this case, it corresponds to the initial co-activation of the two motor-responses, before the activation in the correct response unit inhibits the activity in the other response unit.

These findings indeed match some of the electro-physiological data in the conflict literature, drawing a tight parallel between the ERN and the N2, a negative component observed 200-250 ms after a conflicting stimulus in the flanker, oddball, and go-nogo tasks (Cohen and Yeung, 2006). According to the simulation, the N2 would be a good candidate to explain the occurrence of conflict prior to the response in correct trials. Indeed, it has been shown that the ERN and the N2 have very similar generators in ACC (Van Veen and Carter, 2002) and the link between the prediction of the model and the N2-ERN components seems to be further validated (Yeung et al., 2004; Van Veen and Carter, 2002;

Veen and Carter, 2002; Cohen and Yeung, 2006). Interestingly, the conflict monitoring framework not only provides a model for the occurrence of the ERN but also proposes a model of error detection. Indeed, the detection of conflict corresponding to the evaluation of congruence between the action and the correct response provides a way of detecting errors. These findings suggest that error detection can also be modeled by post- response conflict that signals the occurrence of an error whenever its value crosses a given threshold.

What is the exact difference between the mismatch and the conflict theory? Given that modeling of the conflict theory is much more detailed and the underlying neural mechanisms much more precisely investigated than that of the mismatch theory it seems difficult to compare the two. However, we can see that both theories have in common the fact that errors result from responses that do not make full use of the available information. More importantly, both theories suppose that error detection results from the discrepancy between the motor response and the correct response, which is computed from the evidence accumulated after the motor response. However, some differences remain between the two theories (Yeung et al., 2004) specifically in the dynamics and timing that they suggest. One major distinction is that conflict theory relies on an assessment of conflict in a continuous fashion, the occurrence of a conflict signal not being time-locked to any particular neural event. In contrast, according to the comparison or mismatch theory, the comparison process should be locked to the motor response. Alternatively, some authors have proposed that the mismatch signal reflected in the ERN could be locked to the final computation of the correct response (Falkenstein et al., 2000) but this hypothesis has not been yet carefully tested. In both cases, the mismatch theory supposes that the ERN is strictly time-locked to one event while the conflict theory does not make any such prediction, presenting an important difference between the two models. Another question is which exact information is taken as an input for the ERN, a point that should also have some impact on its dynamics. With regard to that matter, conflict theory hypothesizes that the ERN is driven by the activity still present in the response unit which can be assumed to model the activity in motor cortex, therefore time-locking the ERN to the motor response. Crucially, the incorrect motor response can still be active and trigger the ERN since any simultaneous activity in both response units triggers a conflict signal. This could explain why the ERN seems to start almost simultaneously with the onset of the response. As the mismatch theory is less clear on this point, it is again difficult to determine what its predictions are in this respect. Nonetheless, it can be predicted that if the computation of the correct response is delayed, either the ERN should be reduced in amplitude if it is in fact locked to the motor response or it should be delayed in time as predicted by Falkenstein et al. (2000). Secondly, mathematical views of the two types of models lead to an interesting dissociation: while the comparison model relies on the subtraction of the signal from the actual and the correct motor response, the conflict model proposes that the ERN reflects the product of the two signals. Interestingly, while the two models make very similar predictions on the ERN amplitude when signal regarding the motor and the correct-response are strong, they make rather different predictions when one of the signals is very weak. In particular, the subtraction will produce a relatively strong signal even when subtracted from a near zero value. On the contrary, the product will be very close to zero in this situation, conflict

being virtually absent. Therefore, the two models should be disentangled when signals about the motor or the correct response are weak. Indeed, the existence of a small "default" negative signal when information concerning the correct response is reduced was found in several studies (Pavone et al., 2009; Woodman, 2010; Pailing and Segalowitz, 2004a) and still needs to be addressed by conflict monitoring theory.

A third theory proposed by Holroyd and Coles (2002) tries to place the ERN in the framework of Reinforcement Learning Theory which gives a central role to basal ganglia and their dopaminergic projection in the ACC. According to this theory, the ERN results from the interruption of dopaminergic inhibition on the ACC when a negative reinforcement signal is emitted, i.e. when the consequences of the action are worse than expected. In this framework, the ERN amplitude is influenced by a learning signal carried forward into the cortical generators of the ERN by the mesencephalic dopamine system. This mode explains the presence of the FRN, a negative signal with similar distribution to the ERN, when negative feedback is given. In the absence of feedback however, the ERN reflects the negative reward signal associated with incorrect association of stimulus and response. Importantly however, according to this view, the ERN and its underlying source, the ACC, would not reflect an ongoing monitoring process but rather result from the signal of the basal ganglia indicating a worse-than-expected outcome of the action. In this sense, the ERN would constitute a true prediction error signal, that would play an important role in learning. According to their model (Figure 2.6), the ACC would play the role of a "motor control filter" which would decide which motor command among the ones computed by other controllers is sent to the motor system. Indeed Holroyd and Coles (2002) propose a model of how error-detection signal can be integrated to a more global learning process rather than how errors themselves are detected. The prediction of this theory has been tested (Holroyd et al., 2003; Holroyd and Coles, 2002; Holroyd et al., 2009) and seems to account for some findings regarding the role of the ERN in learning.

In conclusion, the question of the function of the ERN is still the object of an important debate which is yet to come to a consensus.

2.3 Consciousness and the ERN

2.3.1 Variation of the ERN with confidence ratings

One of the most debated questions in recent years is the relationship between the ERN and error awareness. Does the ERN reflect the subjective experience of making an error? How does the ERN vary with certainty about the response and the certainty about the stimulus? In their original paper, Gehring et al. (1993) found that greater ERN amplitude was associated with less strong hand-grip responses, suggesting that responses that were more uncertain were associated with greater ERN. Furthermore, the authors also found that a greater ERN was associated with higher probability of correction, further suggesting that the ERN reflects a form of knowledge concerning the correct response. Since this first article, the question has been systematically investigated and debated (see Wessel, 2012 for a review).

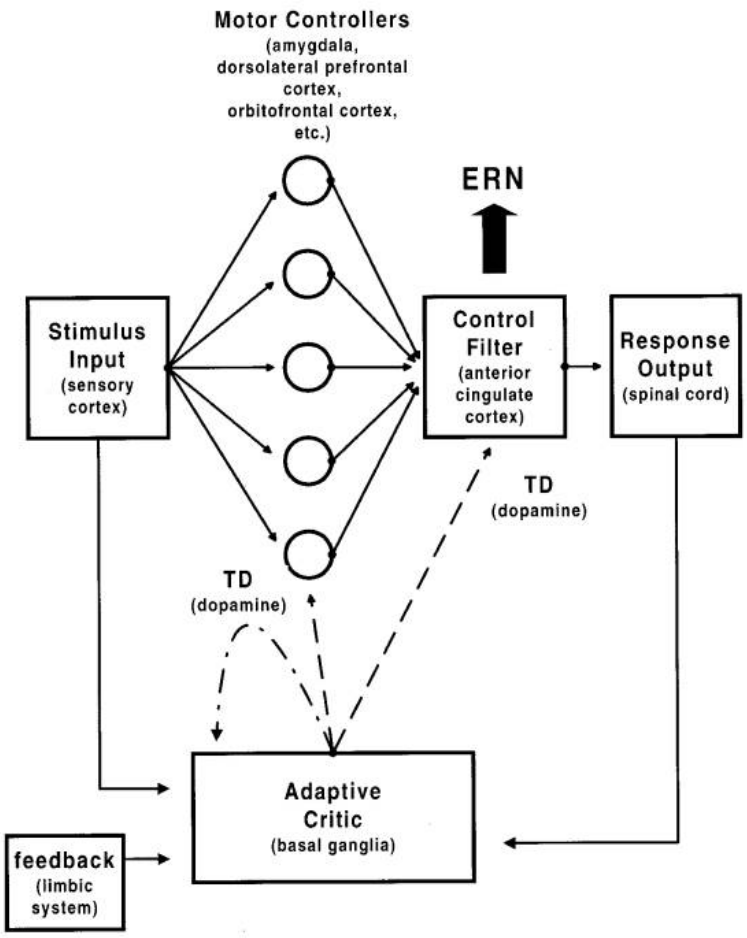


Figure 2.6: Model of conflict monitoring (from Holroyd and Coles, 2002). Each component corresponds to a neural substrate is given in parentheses below each box. See text.

In a seminal paper by Scheffers and Coles (2000), the authors investigated more precisely the question of how the ERN varies with subjective perception of response accuracy. In an Eriksen Flanker task, the authors ask the subjects to rate for each trial the confidence they had in their response on a scale with five levels, ranging from "Sure Correct" to "Sure incorrect". Investigating the level of negativity in each of the five subjective ratings, they observed that the negativity amplitude varied with the subjective confidence reported by the participants, independently of the objective performance. In other words, both correct and error trials were associated with a large negativity when they reported being sure of being incorrect while the negativity was significantly reduced when they reported being sure of being correct (Figure 2.7). As this analysis could be performed only on a smaller number of participants that had enough data in each of the categories, this analysis was confirmed by pooling together the "Don't know" responses of the subjects showing that in this case also, the amplitude of the ERN varied in a linear way with the subjective confidence reported by the subject.

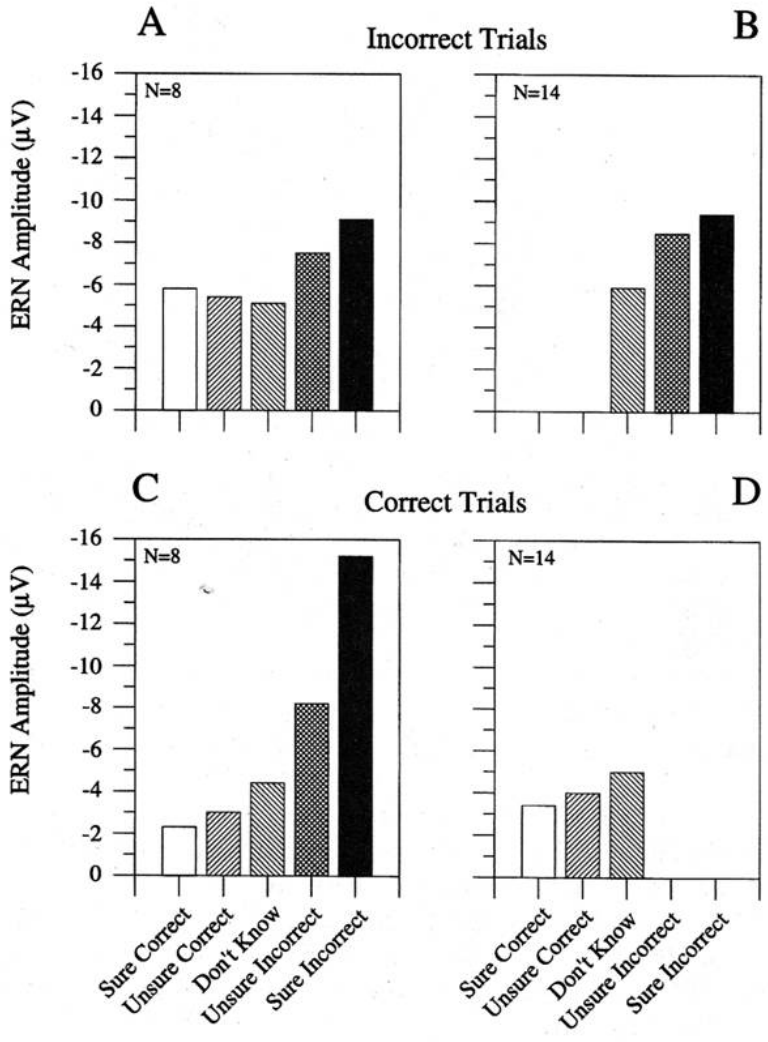


Figure 2.7: Variation of the ERN with subjective confidence (from Scheffers and Coles, 2000). Mean error-related negativity (ERN) amplitude as a function of subjective perception of accuracy for incorrect (top graphs) and correct trials (bottom graphs). Left graphs correspond to the data of 8 subjects who had enough data-points in each of the five subjective ratings. Graph on the right present the data of all fourteen subjects. According to this result, the ERN amplitude varies with the subjective perception of performance.

Following this important finding, other studies have tried to replicate these results. In their recent work using a "digit entering task" in which a sequence of five digit needs to be repeated after a very short presentation duration, Hewig et al. (2011) show that the ERN is indeed modulated by confidence in the response as assessed by a three levels scale. Such results have also been replicated in a recent study using wagering on performance to assess confidence in the response (Shalgi and Deouell, 2012). Indeed, the authors found that wagers had a major impact on the amplitude of the ERN, erroneous trials on which subjects bet with certainty that they were correct being associated with smaller ERN.

2.3.2 The ERN in anti-saccade paradigms.

While this compelling pattern of results suggests that the ERN is tightly correlated to subjective performance rating, the debate was renewed by a striking study by Nieuwenhuis et al. (2001) which showed that the Ne can be observed even after "Unaware errors", i.e. errors that subjects failed to report. This intriguing result was very important as it suggested that firstly, the ERN is not related to conscious error detection and secondly that performance monitoring may occur outside consciousness. Such a finding was particularly noteworthy as it constituted one of the first pieces of evidence that higher-order cognitive functions related to action monitoring can remain perfectly operational without gaining conscious access. Indeed it placed the ERN at the top rank of brain markers of non-conscious processing. Importantly, this result was obtained with a very specific protocol (Nieuwenhuis et al., 2001) based on an occulomotor task where the subject had to make a saccade in the opposite direction of a cue (anti-saccade task). Crucially, in this task the subjects had to inhibit their spontaneous eye movements in the direction of the cue to make a correct response leading the subjects to make a lot of errors: in many trials, they initiated a saccade in the cue direction and then corrected it by making a correct saccade in the opposite direction. A crucial result of this experiment was that subjects failed to consciously report making errors for a significant number of trials where their initial movement was incorrect, leading to a mixed pattern of aware and unaware errors. While the ERN remained present for these partial error trials, even when subjects did not detect their initial erroneous movement, only the later Pe component varied with the subjective error awareness of the subjects. These interesting results were further replicated (see Figure 2.8) more recently by Endrass et al. (2007); Endrass et al. (2005); Endrass et al. (2012) in a similar anti-saccade task. The authors obtained identical results with the only difference being that the late (300-400 ms) but not the early part (200-300 ms)of the Pe component was related to conscious error detection.

2.3.3 An ERN for undetected errors

An important confound of the anti-saccade paradigm is that almost all the trials that were considered as unaware errors were in fact followed by a quick correction saccade. This fact may explain why subjects failed to categorize these trials as erroneous. Moreover it suggests that the modulation of the Pe could be linked to error correction and not to error awareness as proposed by the authors. However, several studies have replicated these results in other various task sets (Wessel, 2012). Using paradigms specifically manipulating error awareness by confusing instructions or task settings, several studies found that the ERN indeed remained present independently of whether the subjects were aware or not of their errors (O'Connell et al., 2007; Dhar et al., 2011; O'Connell et al., 2009; Shalgi et al., 2009). Importantly, they found similar results in favor of the hypothesis that the Pe was linked to awareness of the error in a very different paradigm (O'Connell et al., 2007). These findings were further supported by studies using fMRI and showing identical brain activity in ACC for aware and unaware errors (Klein et al., 2007a; Hester et al., 2005) In contradiction with these results and in

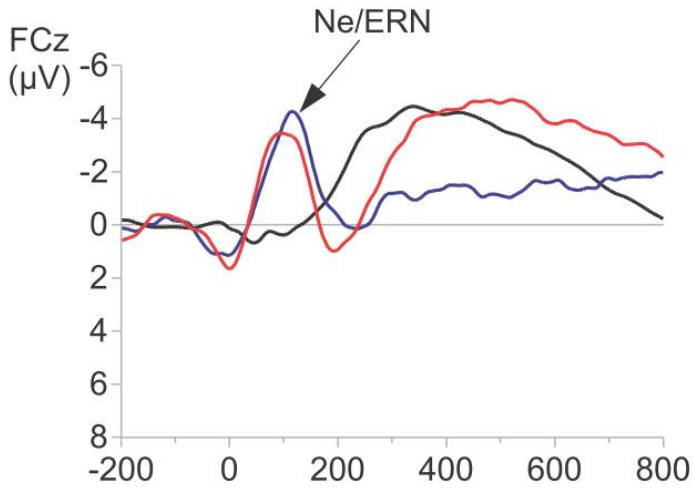


Figure 2.8: The Error-related Negativity is present even when subjects remain unaware of their errors (from Endrass et al., 2007). Graphs depict the grand-averaged event-related potentials (ERPs) elicited by aware errors, unaware errors and correct responses at FCz (baseline: -200 to -100 ms).

accordance with initial results obtained by Scheffers and Coles (2000) however, other studies found that the ERN varied with subjective report of error awareness with a reduced ERN being observed in unaware errors (Hewig et al., 2011; Maier et al., 2008; Wessel et al., 2011; Shalgi and Deouell, 2012; Steinhauser and Yeung, 2010). Indeed, when re-analysing some of their findings, (Orr and Carrasco, 2011) found that the error-related dorsal ACC activity was significantly greater during aware errors supporting the possibility that their initial null result was due to low statistical power.

Nevertheless, while all these results seem to argue that the ERN is related to subjective report of confidence, converging evidence suggests that it constitutes a relatively automatic process which is not sufficient to lead to awareness of the error by itself, whereas the Pe seems to be very directly linked to the conscious experience of making an error and its subsequent signaling. This result on the ERN is coherent with studies by Rabbitt et al on error correction (Rabbitt, 2002) showing that even errors that are neither reported nor recalled are registered at some level as they are followed by slower trials.

2.3.4 An ERN in subliminal condition?

A related but quite different question however, is whether an ERN can be evoked by subliminal stimuli, when performing a task on masked images. A few studies have investigated this question and obtained mixed results.

A first study that addressed this issue (Pavone et al., 2009) claimed to observe a significant ERN on both unaware and aware errors compared to correct trials. However, the protocol used made it difficult to interpret the results: the subjects were presented either with unilateral or bilateral checkerboards, one of them being made barely visible by adjusting its luminance to threshold level. Subjects had to indicate

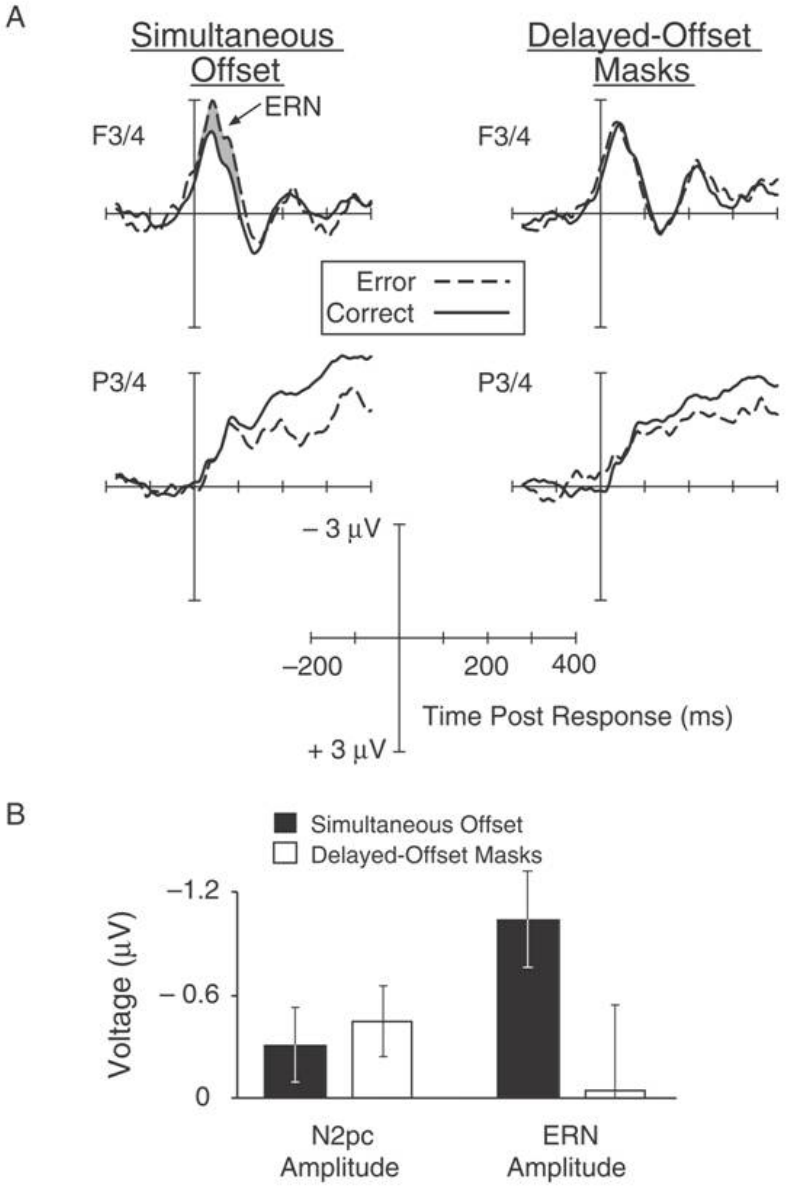


Figure 2.9: The ERN is absent in masked conditions from Woodman, 2010. A: The graphs depict the grand average across participants of the ERPs time-locked to the motor response for unmasked (simultaneous-offset, left) and masked (delayed-offset trials right) stimuli, in correct (solid lines) and incorrect responses (dashed lines).The ERN (shaded surface) is observed only in the masked condition. B: The bar-plots depict the amplitude of the N2pc and the ERN for unmasked (simultaneous-offset) and masked (delayed-offset) trials.

whether they saw one or two checkerboards and then indicate if they had made a mistake or not. Therefore, the error-detection task operated on the detection task itself. While a significant difference was observed between unaware errors and correct trials, their results were difficult to interpret. In particular, correct trials corresponded to several types of trials: trials in which only one checkerboard was presented

(easy trials) and trials in which two checkerboards were presented at very different luminance (hard trials). As the results for each type of trials were not reported and difficulty is known to modulate the CRN- the negativity seen in correct trials- it is difficult to determine the true amplitude of the differences observed. This is particularly problematic as a close examination of the graphs shows very little difference between error and correct trials, regardless of error awareness, as well as important pre-response baseline variations suggesting that the amplitude of the negativity is similar when baseline correction is applied. A more strictly controlled study was published one year later as a re-analysis of previously published data (Woodman, 2010). Using 4-dot masking, Woodman (2010) observed an ERN when the target was consciously perceived, but not when it was masked and became invisible (Figure 2.9). Importantly, in this paradigm consciousness of the stimulus was not assessed by subjective report but by the masking condition, making difficult to evaluate how such manipulation affected subjective perception of the target. Interestingly however, Woodman (2010) found that the N2pc component that preceded the ERN remained present even during the masked condition, suggesting that this form of masking might not drastically diminish available information on the stimulus. Therefore this study suggests that indeed the ERN is absent in subliminal conditions. Finally a more recent study (Hughes and Yeung, 2011) found that the ERN was indeed reduced when the stimulus perception was degraded by masking. While this study did not assess awareness of the stimulus on a trial by trial basis, it nonetheless suggests that consciousness of the stimulus has an impact on the amplitude of the subsequent ERN.

Therefore, while the results concerning the relation between awareness and the ERN appear unclear when considered in their globality, they form a more coherent pattern when we consider which aspect of awareness was manipulated in each study. We tried to gather studies according to which factor was affected in each study: stimulus awareness, action awareness or error awareness. The pattern of results (Table 2.1) suggests that:

1. The ERN is present when the action is unaware
2. The ERN is absent when the stimulus on which the task is performed is subliminal
3. The presence of an ERN itself does not imply awareness of making an error

2.4 Schizophrenia, Metacognition and Consciousness

2.4.1 Psychopathology and the ERN

The ERN is known to be abnormal in many pathologies and its variations are predictive of several abnormal phenotypes. In particular, the ERN amplitude varies with age: in older adults, studies have shown that the ERN is decreased in amplitude compared to young adults (Falkenstein et al., 2001; Mathalon et al., 2003; Nieuwenhuis et al., 2002) though these results have sometimes being criticized on the grounds that performance in older adults is often decreased compared to younger subjects (Olvet and Hajcak, 2008).

Name	Year	Task	Subjective signalling	N	*p-value*	Statistical Test
Effect of confidence Conscious stimulus / Conscious Action						
Scheffers and Coles (all)	2000	Flanker task (letter version)	"Five-point scale ranging from "surely incorrect" to "surely correct"	8	0.005	ANOVA (two-sided)
Scheffers and Coles (partial)	2000		"Three-point scale ranging from "Don't know" to "surely correct"	15	0.002	
O'Connell et al.	2007	Manual Go-NoGo Task, visual stimuli	Awareness button on next trial, abolish Go response	12	0.872	ANOVA(two-sided)
Maier et al.	2008	Flanker task (letterversion) with additional neutral stimuli	Awareness button (1200 ms time including primary task)	14	< 0.001	ANOVA(two-sided)
Shalgi et al.	2009	Manual Go-NoGo Task, auditory stimuli	Awareness button on next trial, abolish Go response	16	0.187	t-test(two-sided)
Steinhauser and Yeung	2010	Visual pattern discrimination	Awareness button (1000 ms time)	16	0.046	t-test (two-sided)
Dhar et al.	2011	Manual Go-NoGo Task, visual stimuli	Awareness button (1500 ms time)	14	0.467	t-test (two-sided)
Shalgi et al.	2012	Manual Go-NoGo, visual shapes	Wagering	12	<0.01	t-test (two-sided)
ERN for Non-conscious action						
Nieuwenhuis	2001	Anti-saccade task	Awareness button (1250 ms time)	15	<0.001	ANOVA (two-sided)
Endrass et al.	2005	Oculomotor stop-signal task	Binary rating (1300 ms time)	20	N.A.	ANOVA (two-sided)
Endrass et al.	2007	Anti-saccade task	Binary rating with an "unsure" option (press both buttons)	19	<0.001	t-test(two-sided)
Wessel et al. (Exp. 1)	2011	Anti-saccade task	Binary rating	17	0.027	ANOVA
Wessel et al. (Exp. 2)	2011	Anti-saccade task	Binary rating (with post-hoc "sureness" quantification based on rating times)	17	0.018	ANOVA
ERN for Non-conscious stimulus						
Pailing et al.	2004	Dual task, divided attention (letter discrimination and auditory judgement)	No subjective judgement	13	>0.05	ANOVA
Pavone et al.	2009	Visual pattern discrimination (low luminance)	binary rating	10	0.044	t-test (two-sided)
Woodman	2010	Visual search with non-masked and masked stimuli N2pc	No subjective judgement	7	>0.05	ANOVA(two-sided)
Hughes and Yeung	2011	Flanker task (arrow version) with additional masked stimuli	Awareness button (1000 ms time)	8	<0.001	ANOVA on main effect of performance
Hewig et al.	2011	Semi-blind digit-entering	"Three-point scale ranging from ?surely incorrect? to ""surely correct"""	16	>0.05	ANOVA

Table 2.1: ERN and consciousness, review of the different articles (adapted and corrected from Wessel (2012))

Altered ERN has been associated with several pathologies including anxiety and depression. The ERN of those with obsessive-compulsive disorder (OCD), has been reported to be increased when compared with age matched controls (Gehring et al., 2000). Furthermore, it has been shown that children suffering from generalized-anxiety disorder (GAD) also possess an increased ERN (Weinberg et al., 2010; Johannes et al., 2001; Ladouceur et al., 2006). This finding has been associated with results showing overall hyper-activity in ACC for OCD patients (Fitzgerald et al., 2005). Interestingly, it was shown that even after a treatment that was successful, the ERN remained increased in OCD children suggesting that it might constitute a trait-like marker for the pathology (Hajcak et al., 2008). However, the ERN was not found to vary significantly with the state of anxiety: when anxiety was induced in spider phobic subjects, their ERN did not increase compared to when no stressful stimuli were presented (Moser et al., 2005), suggesting that the ERN was not simply modulated by state of anxiety.

Abnormal ERN has also been shown in depressed individuals. The ERN is increased compared to controls in subjects suffering from depression in various task sets (Holmes and Pizzagalli, 2008; Chiu and Deldin, 2007) and was shown to be associated with prediction of recovery from depressive symptoms in elders (Kalayam and Alexopoulos, 2003). One study found that the ERN was increased for negative but not for positive rewards suggesting that the ERN might be specifically modulated by negative outcomes in depressed individuals (Chiu and Deldin, 2007). This finding is coherent with other results showing that indeed part of the cingulate cortex, in particular its most rostral part, was abnormally active in depressed patients (Steele et al., 2004). These findings led some authors to propose that the increased ERN in depression and anxiety might not be specific to these pathologies, but rather reflect a common abnormal mechanism linked to negative affect (Hajcak et al., 2004; Olvet and Hajcak, 2008; Hajcak and Foti, 2008) that translates into an increased sensitivity to committing errors. Interestingly, the ERN was also found to be decreased in other pathologies. For example, the ERN of both non-medicated and medicated patients suffering from Parkinson disease were attenuated compared to those of healthy controls matched for age (Stemmer et al., 2007).

With regard to ERN psychopathology, it is interesting is to determine how the ERN varies with lesions in the prefrontal cortex. Unsurprisingly, the ERN was found to be attenuated in patients suffering from orbito-frontal lesions while doing a manual stroop task (Turken and Swick, 2008). Interestingly however, while performance in error correction seemed to be affected in several of these patients, post-error slowing was found to be impaired in all the patients, a finding that might validate the special role in cognitive control mechanisms of post-error slowing (Logan and Crump, 2010).

As focal lesions in cingulate are quite rare only a few studies have reported the results of lesion in this area and their impact on response to errors. However, a single case study on patient R.N. suffering from a very focal lesion in right ACC (Figure 2.10) showed not only that the post-response negativity was indeed present and peaking at the exact same time as the ERN but also that it was present on both errors and correct responses, showing that the ERN was attenuated and not different in amplitude from the CRN (Swick et al., 2002). Interestingly however, the N2 remained preserved in this patient responding normally to conflict, a result that speaks in favor of a distinct neural source for the N2 and the ERN.

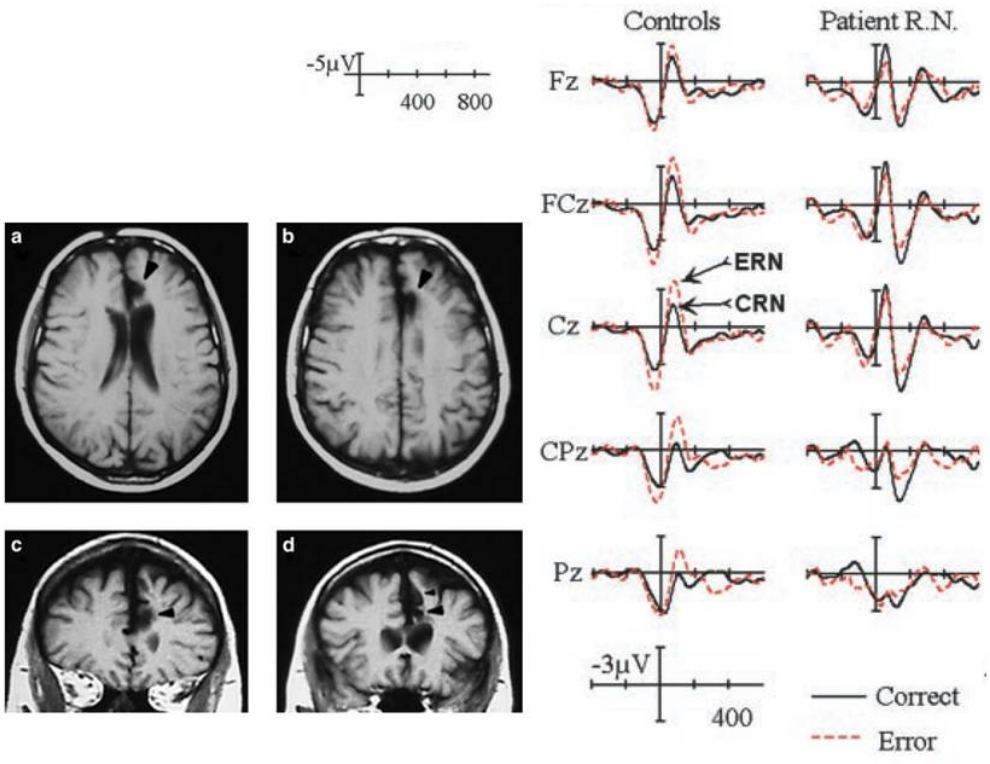


Figure 2.10: The ERP results in response to error from patient R.N. suffering from a focal lesion to ACC (from Swick et al., 2002). Horizontal sections of MRI scans illustrate the lesion in the left ACC (indicated by black arrows). Graphs below show the response-locked ERN and the CRN from frontal (top) to parietal (bottom) electrodes. Negative is plotted upward.

Nonetheless, these findings should be treated with caution: as the region of the cingulate was impaired only unilaterally, it is difficult to know if the recorded response corresponded to a partial response of the preserved cingulate zone (Ullsperger, 2006).

Inter-individual ERN variation has also been investigated in association with genetic markers. In particular, the ERN amplitude was correlated with allelic variant of gene 5-HTTLPR that controls region of the serotonin transporter (5-HTT) and which has been shown to be associated with depression. Individuals carrying short variants of the allele presented significantly higher ERN amplitude than age- and gender-matched individuals homozygous for the long allele (Fallgatter et al., 2004). While these results needs to be treated with caution, they nonetheless speak in favor of the ERN as an important index of normal or impaired cognitive control functioning.

2.4.2 Error detection and the ERN in schizophrenia

Data from the literature suggest that some processes related to error detection are altered in schizophrenia. In particular, several studies show that the ERN is attenuated in this population (Kerns et al., 2005; Foti et al., 2012; Bates et al., 2004; Bates et al., 2002; Mathalon, 2002; Kopp and Rist, 1999;

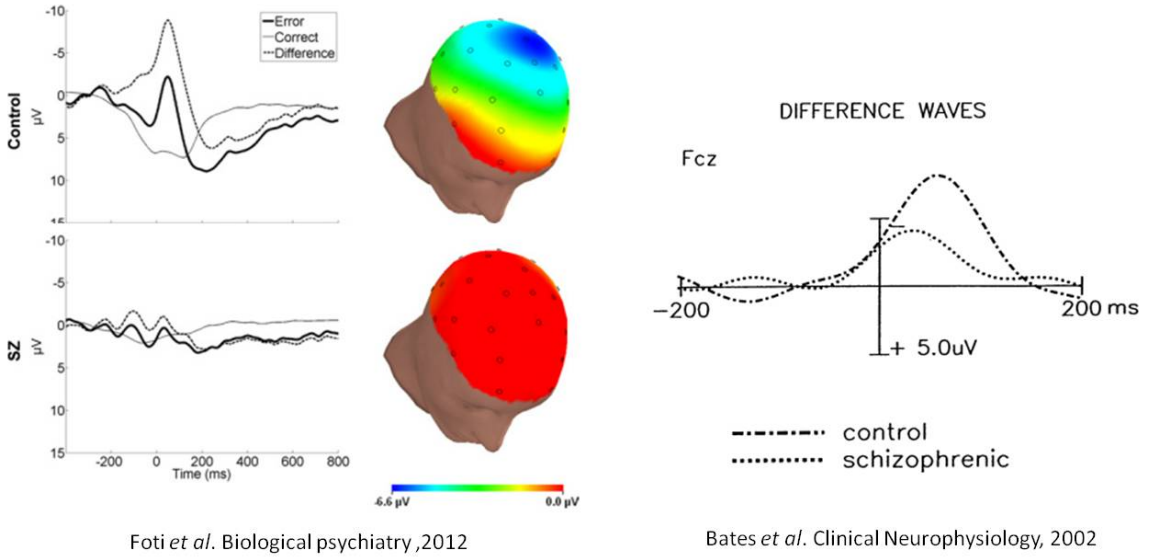


Figure 2.11: Reduced or absent ERN for schizophrenic patients (from Bates et al., 2002; Foti et al., 2012). On the left, the graphs depict the waveforms for error (black line) and correct trials (grey line) for control subjects (top) and schizophrenic patients (bottom) at channel Cz. The EEG topographies show the difference between error and correct trials from 0 to 100 msec. On the right, the graphs depict the grand-average plots at Fcz for the difference between error and correct trials for control subjects (dashed-dotted line) and the schizophrenic patients (dotted line).

Morris et al., 2006; Morris et al., 2011; Alain, 2002; Kim et al., 2006; Olvet and Hajcak, 2008; Carter et al., 2001; Laurens, 2003; Hajcak et al., 2004; Pailing and Segalowitz, 2004b). Interestingly, the difference in ERN amplitude has not been restrained to electro-physiological evidence from errors. Indeed, several studies found that schizophrenic patients also present a larger CRN amplitude, comparable in magnitude to their ERN (Mathalon, 2002; Morris et al., 2006; Kim et al., 2006), similarly to what has been found in patients with prefrontal lesions (Swick et al., 2002). In particular, this reduction in ERN amplitude was present even when maximizing the ERN amplitude by emphasizing accuracy over speed. However, functional characteristics of the ERN appeared to be intact as CRN increased when emphasizing speed over accuracy, as has been reported in healthy subjects (Morris et al., 2006).

An interesting distinction has been made among different patient groups (Foti et al., 2012). In a very well controlled study, using an adequately matched population, it was shown that while both patients with schizophrenia and other psychoses presented impaired ERN, the Pe however was impaired only among individuals with schizophrenia, indicating a different relationship to psychotic illness of the two components. Interestingly, the ERN was also associated with more severe negative symptoms (Foti et al., 2012).

Deficit of activity in prefrontal cortex has been documented in schizophrenia and constitutes a possible source of the impairment observed in the pathology. In fMRI studies, it has been shown that schizophrenic patients present altered responses to errors, with decreased activity in ACC (Carter et

al., 2001) following incorrect responses compared to normal subjects. Similarly, Laurens (2003) found that activity in the rostral ACC was specifically reduced in individuals with schizophrenia compared to age-matched healthy controls when committing errors. Additionally, studies showed that anatomical differences in ACC could be observed in schizophrenic patients (Zetzsche et al., 2007), in particular in rostral ACC regions of right hemisphere. Processes related to cognitive control seem to be globally altered in schizophrenic patients. In particular, several studies suggest a specific impairment of cognitive function associated with proactive cognitive control tasks (Barch et al., 2001). Proactive control is described as the part of the cognitive control processes that are related to the early activation of goal-relevant information, which is maintained in an anticipatory manner for further tasks. In that respect, this is the form of control that may be linked in the closest manner to consciousness, being involved in orienting attention, perception, and action systems to a particular conscious content. Schizophrenic patients indeed present strong deficits (Lesh et al., 2013) in prefrontal activity and in particular in dorsolateral prefrontal cortex (DLPFC) in proactive compared to reactive control tasks. Dysfunction of the prefrontal cortex may explain these findings (Barch and Ceaser, 2011) since these deficits have been present at an early stage of the disease, prior to the administration of medication (Barch et al., 2001). This impairment could be linked to deficits in working memory that has been widely shown in patients. In particular, an interesting study showed that for the same perceptual information, the ability to retrieve information from working memory was specifically impaired in patients (Smith et al., 2011), explaining deficits in other associated functions.

2.4.3 Schizophrenia and Consciousness

Several studies suggest that non-conscious brain functions remain fully functional in schizophrenia. A study demonstrated that implicit learning was not impaired in schizophrenic patients compared to a control population (Danion et al., 2001). In particular, studies performed by colleagues using direct masking paradigms (Dehaene et al., 2003; Del Cul et al., 2006) in order to study the treatment of subliminal information in schizophrenic patients showed that it was left unimpaired in these patients.

In one study (Dehaene et al., 2003), it was demonstrated that in a task causing a conflict between two contradictory responses, patients have decreased brain markers of conflict only in conscious trials, while classical effects in subliminal conditions were preserved. Schizophrenic patients and control subjects matched in age and years of study performed a number comparison task in which a prime number preceded the target in a fully visible or in a masked manner. Interestingly, their results showed that while the prime elicited reliable repetition priming in both schizophrenic patients and controls (Figure 2.12, left panel), conscious conflict evoked by incongruent primes was strongly reduced in patients, as was related brain-activity in ACC region (Figure 2.12, right panel). This result showed a specific impairment in conscious conflict monitoring for patients, suggesting a specific deficit in schizophrenia of conscious activity in ACC.

A second study (Del Cul et al., 2006) investigated more carefully the conscious access of

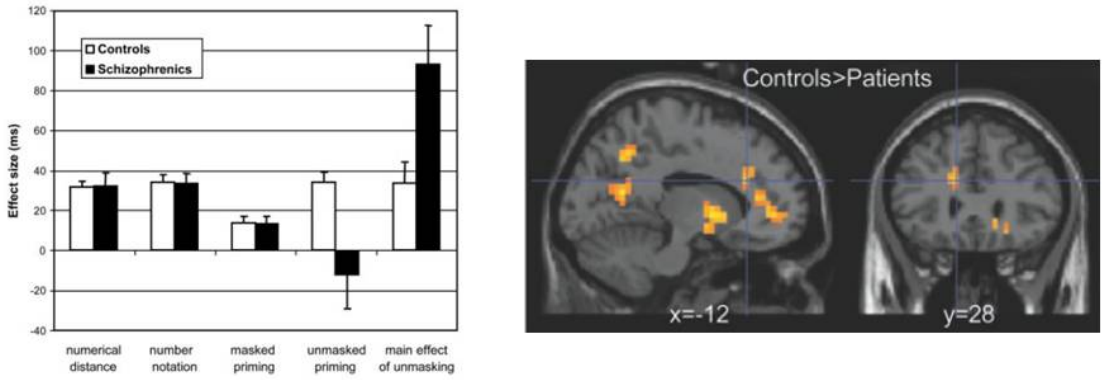


Figure 2.12: Preserved subliminal priming but impaired conscious conflict-related activity in ACC for schizophrenic patients (from Dehaene et al., 2003). Left graphs depicts the priming effect in control subjects (white bars) and in schizophrenic patients (black bars) according to different experimental conditions. In particular, effect of subliminal priming was identical in both groups (middle bars). Right graphs show the brain regions in which the conflict * visibility interaction was stronger in control subjects than in schizophrenic patients. In particular, this analysis revealed greater conflict-related activity in ACC for controls than for patients.

schizophrenic patients in a masking study. The results confirmed those previously obtained (Green et al., 1999; Saccuzzo et al., 1996), showing that schizophrenic patients' threshold for access to consciousness in backward masking paradigm was higher than those of controls (Figure 2.13, right panel). This increase of the threshold of consciousness appears to correlate with schizophrenic symptoms (positive and negative symptoms, as well as disorganization). In addition, approximately 30% of patients presented a phenomenon of "visual illusions" and reported seeing stimuli that did not correspond to the ones that were presented. Interestingly however, schizophrenic patients presented preserved non-conscious response to visual stimuli as measured by subliminal priming (Figure 2.13, left panel). These findings suggest that there might be a specific alteration of processes related to access to consciousness in schizophrenia and that this alteration may be related to a disturbance in late stages of stimulus processing while the non-conscious early stages remain preserved.

This hypothesis is consistent with results showing that schizophrenia is associated with functional disturbance of large-scale integration processes caused by abnormal long-distance cortico-cortical and cortico subcortical connections (Friston and Frith, 1995; Friston, 1998; Friston, 2005; Haraldsson, 2004; Liang et al., 2006; Schmitt et al., 2011), in particular in prefrontal cortex (Fletcher et al., 1999; Grillon et al., 2012). These abnormalities in connectivity may cause deficits in the temporal integration of information between distant brain regions (Uhlhaas et al., 2008). For some authors, this impaired connectivity would result in the disruption of processes that require integration of high-level information (Bassett et al., 2008), contrasting with the preservation of some more automatic functions. These results are consistent with the predictions of the global neuronal workspace model (Dehaene and Changeux, 2011) which predicts that conscious access is based on the long distance connections between remote

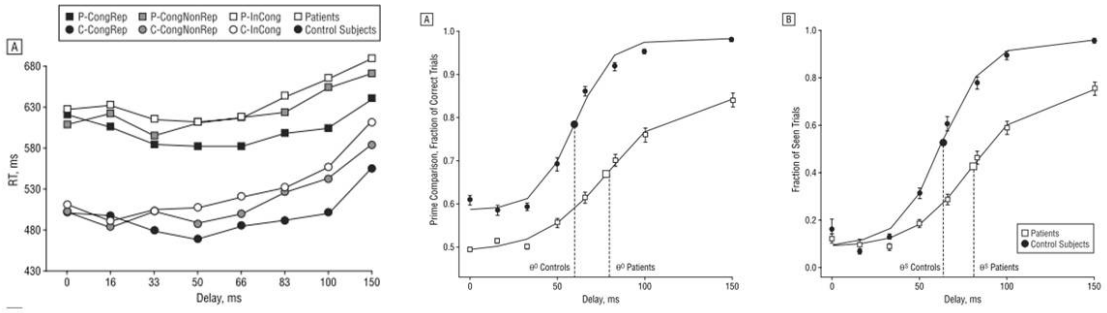


Figure 2.13: Preserved subliminal priming but impaired conscious access for schizophrenic patients (from Del Cul et al., 2006). Left graph measures the priming effect during the target number comparison task, plotting the mean reaction time (RT) for each group, each condition of prime-target relation, and each delay. Response priming was defined as the difference in reaction time between incongruent (InCong) and congruent non-repeated (CongNonRep) trials and repetition priming as the difference between CongNonRep and congruent repeated (CongRep) trials. Right graphs show the objective and subjective measures of access to consciousness as the percentage of correct responses in the prime categorization and the proportion of trials subjectively rated as "seen" as a function of prime-target delay . In both graphs, black lines correspond to the sigmoid fit of the data.

brain areas.

CHAPTER 3

An experimental approach to study consciousness and metacognition

---

In the first chapter of this manuscript, we saw that several questions concerning the relationship between consciousness and metacognition remain unanswered. In particular, it has been proposed that metacognitive knowledge is tightly linked to consciousness (Kolb and Braun, 1995; Rounis et al., 2010; Lau and Passingham, 2006). Moreover, measures of consciousness relying solely on metacognitive knowledge have been proposed (Persaud et al., 2007). However, evidence that some metacognitive processes may occur outside of consciousness has been shown. Many cognitive control mechanisms are known to be triggered in subliminal conditions (Cohen et al., 2009; van Gaal et al., 2008; van Gaal et al., 2009; Lau and Passingham, 2007; Pessiglione et al., 2008; Pessiglione et al., 2007). Furthermore, some performance monitoring systems appear to be triggered non-consciously with errors being detected by the brain while subjects remain unaware of making them (Nieuwenhuis et al., 2001; Endrass et al., 2007; Logan and Crump, 2010; Cohen et al., 2009).

Therefore, the link between metacognition and consciousness remains unclear. In the present work, we tried to shed some light on this question by testing systematically how visual awareness influenced further cognitive processes related to action selection and performance monitoring. We focused on error-detection as a simple and yet crucial metacognitive task. Our goal was to address several key questions that we believe remain to be answered concerning the relationship between consciousness and metacognition:

- Can information about performance be extracted non-consciously?

  While some indirect measures seem to indicate that performance monitoring systems can be triggered non-consciously (Logan and Crump, 2010), very few studies have (Kanai et al., 2010) directly investigated how subjects perform in a forced-choice error detection task when responding to subliminal stimuli. What information are subjects able to report on their performance in subliminal conditions?

- Can brain signals related to performance monitoring be evoked in subliminal conditions?

  While several studies confirmed that the ERN may be present when errors are not detected consciously, unclear results have been obtained in true subliminal conditions where the stimulus on which subjects performed the task was presented non-consciously (Woodman, 2010;

Pavone et al., 2009). More importantly, they did not allow the determination of how subjective perception alone, above variation in masking strength, influences the amplitude of the ERN.

- Can we find markers in response selection and action monitoring of the crossing of the threshold for conscious access?

  While it has been shown that non-conscious stimuli are processed by the brain and activate a series of specialized cognitive modules up to high computational stages (Naccache et al., 2005; Van den Bussche et al., 2009; Sklar et al., 2012; Batterink and Neville, 2013; Pessiglione et al., 2008; Pessiglione et al., 2007), the mechanisms of action selection in non-conscious conditions compared to conscious conditions are still unclear. How does conscious access impact the decision process, action selection and action monitoring? Is it possible to find markers of this process that are modulated solely by subjective visibility?

- What computational models may account for first-order decision, error detection and conscious access?

  Several models of decision and meta-decision have been proposed (Pleskac and Busemeyer, 2010). In particular, different computational models have been developed to account for error detection mechanisms occurring in the brain (Yeung et al., 2004; Falkenstein et al., 2000). In parallel, cognitive models of decision making in conscious and in non-conscious situation have been suggested (Del Cul et al., 2009). However, these models have not been confronted. Is it possible to find a single cognitive model that can integrate these different aspects?

In the following chapter, we are going to present in greater detail the paradigms and methods that we propose to use in addressing these questions.

3.1 Masking study

We have seen that different measures have been proposed, related to distinct cognitive models of consciousness. In the present work, we wanted to assess how subjective visibility alone influences processes related to performance monitoring and error detection. Therefore, we used a masking paradigm developed in the lab (Del Cul et al., 2007; Del Cul et al., 2006; Del Cul et al., 2009) which allows the study in a precise manner of conscious and non-conscious processes (Del Cul et al., 2007). This paradigm uses a subjective measure of consciousness while also enabling a more objective assessment of perceptual process. In this paradigm, a target number is presented at one out of four positions on a screen and followed by a mask composed of an array of letters. The target number which can be either 1, 4, 6 or 9, is presented for 16 ms while the mask is presented for a longer duration of 250 ms. Importantly, the mask is presented at a variable delay following the offset of the target, the SOA ranging from 16 to 100 ms. In one sixth of the remaining trials, the mask is presented alone (mask-only condition) as a control condition.

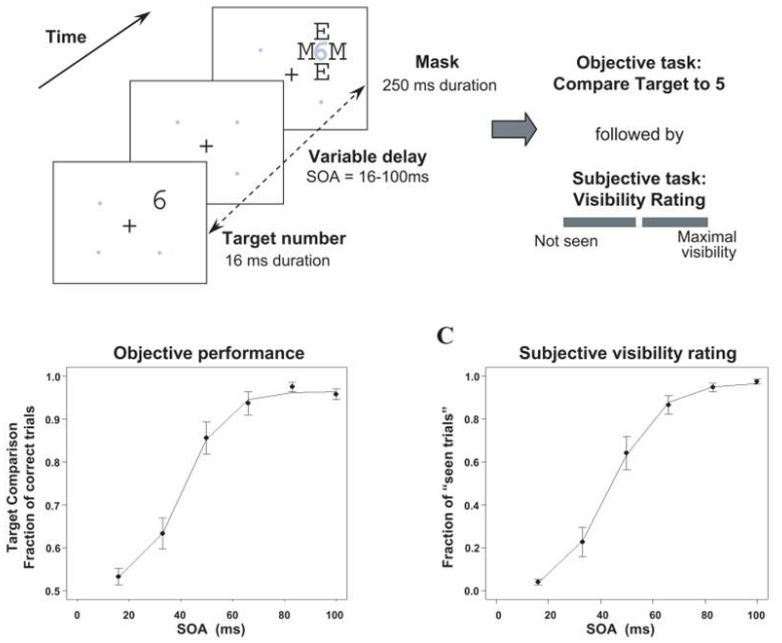


Figure 3.1: Masking Paradigm (from Del Cul et al., 2007). Top image depicts the masking paradigm used. A target number was presented for a short duration (16 ms) at one out of four possible locations. A mask composed of four letters (two E and two M) was presented at the same location but for a longer duration (250 ms), following a variable delay (SOA). Subjects' task was to determine if the number was smaller or bigger than 5 and then rate the subjective visibility of the number using a continuous scale. The graphs below show the percentage correct in the number comparison task and the proportion of "seen" trials as a function of SOA. Both measures increased in a non-linear way with SOA.

Importantly, the subjects were asked to perform two tasks on the masked stimulus. Firstly they had to indicate whether the target number was smaller or larger than five. Secondly, they had to evaluate the subjective visibility of the target on a continuous scale ranging from "not seen" to "maximal visibility". As can be seen in Figure 3.1, both objective performance and subjective measure of the proportion of "seen" trials increased in a non-linear fashion with SOA. This non-linear increase in visibility is thought to reflect the non-linear transition from non-conscious to conscious perception, characteristic of conscious access. Indeed, using this paradigm Del Cul et al. (2007) showed that some components of electro-physiological response to the stimulus followed this non-linear profile, suggesting a possible link between these components and conscious access. In particular, while some early evoked-responses remained unaffected by masking, late ERPs varied with SOA following the same sigmoidal shape as the visibility report. This pattern of activity coincided with the sudden activation of a distributed bilateral fronto-parieto-temporal network around 270 ms after stimulus onset, suggesting this all-or-none onset reflected the ignition of conscious processing.

One important advantage of such a paradigm is that variations of SOA create different degrees of visibility. As can be seen in Figure 3.1 subjective visibility for shorter SOA is close to 0 and performance is at chance while for the longest SOA values, visibility and objective performance are close to ceil-

ing. This paradigm therefore provides both conscious and non-conscious trials ranging from complete subliminal perception to fully conscious perception. Importantly, as subjective visibility progressively increases with SOA, it is possible to find intermediate SOA trials that are reported as fully seen and trials for which the target remains unseen. Therefore, by sorting trials according to visibility report and SOA, this paradigm allows the study of the variation of brain activity induced solely by subjective reports, above the objective conditions of stimulation. Importantly, behaviour can also be quantified from an objective point of view. In particular, both performance in the number comparison task and visibility reports can be quantified using signal detection theory, in order to obtain a clear idea of the perceptual sensitivity and the bias in responding for these two tasks. To do so, responses to the objective task are transformed into hits and false-alarms, choosing arbitrarily one condition as equivalent to signal presence and the other as equivalent to signal absence. Moreover, "seen" and "unseen" reports can also be considered as reports of target presence or target absence. Therefore for each SOA condition, the hit rate, corresponding to the number of time the presence of the target was indeed detected can be compared to the false-alarm rate constituted by the number of "seen" responses in the mask-only condition.

For the present experiment, small modifications were made to the protocol. In particular, as the goal was to study errors occurring in conscious and non-conscious conditions, we added a strong time-pressure to the number-comparison task so that subjects would make a lot of errors even when they fully perceived the target stimulus. Importantly, this time-pressure was imposed for the first response of the number comparison but not to the visibility question, so that reports of visibility would be as accurate as possible. Additionally, we asked the subjects to rate their performance on each trial by providing a binary response "Error" or "Correct" in order to determine their awareness of their own accuracy. This measure also gave us an indication of the sensitivity in processing the target stimulus, in addition to the simple visibility reports.

3.2 M/EEG, a powerful tool to study brain activity

In the present experimental approach, our aim was to study the precise dynamics of stimulus processing, from perceptual stages to action selection and performance monitoring. Therefore, a crucial aspect of our work was to use a neuroimaging device that allows for an excellent temporal resolution. For these reasons, we recorded brain activity with both magneto- and electroencephalography techniques. In the following section, we briefly describe the potential advantage of these techniques in light of their technical details.

3.2.1 A brief description of MEG and EEG techniques

The study of the electromagnetic field in biology is an ancient one. More than 200 years ago, Luigi Galvani studied "animal electricity" showing that when stimulating electrically the leg muscles of a dead frog, it was possible to observe contraction movements. The discovery of action potentials in the middle

of 19th century confirmed the central role of electro-physiology in biology and medicine. Electroencephalography (EEG) and magnetoencepahlography (MEG) record the electromagnetic activity caused by the brain in a non-invasive manner. While EEG recordings have been performed on humans for almost a decade, with the invention by Hans Berger in 1924 of the electroencephalogram, MEG was developed less than 50 years ago. This is due to the greater difficulty in recording very small values of magnetic activities, as well as the much greater cost of the machine.

Briefly, electroencephalographic recordings are obtained by placing electrodes on the head of the subject. Importantly, a conductive substance such as a gel or a paste must be used to create a connection between electric signal recorded on the scalp and the electrode. EEG signals correspond to the difference of potentials between two electrodes: the electrode placed on the scalp and the reference electrode. A ground electrode is also needed in order to obtain differential voltage, subtracting the same voltage value to the scalp-electrode and the reference. In order for the EEG signal, which is of the order of a few microvolts, to be properly recorded and digitized, it needs to be amplified. Furthermore, elements that might be responsible for a decreased signal, such as high impedance of the recording electrode, must be avoided. Additionally, it is important to control for the presence of artefacts, such as muscular activity or external electronic noise. In particular, electro-ocular activity is a common disturbance in EEG signals and is often the object of a separate recording, for further de-noising. Apart from these technical issues, EEG recording still constitutes a simple, cheap and efficient measure of brain activity. In particular, the easy mobility of the system makes it a key method in medical and research studies.

MEG recordings, on the other hand, have proven to be much more difficult technically. The recording of magnetic signal from brain activity was made possible by the development of the superconducting quantum interference devices (SQUIDs) at the Massachusetts Institute of Technology. SQUIDs are a form of particularly sensitive magnetometers based on superconducting loops that can measure extremely weak signals such as those produced by brain activity. It is combined with magnetically shielded rooms that enable removing external static or low-frequency magnetic fields. In ideal conditions, it is then possible to record signal on the scale of the femto-tesla (10^{-15} T). In recent MEG systems, two types of sensors are available: magnetometers which record the "raw" magnetic field, and gradiometers which record the gradient of the magnetic field in one particular direction of space. Importantly, gradiometers are sensitive to specific spatial patterns of magnetic field and therefore record almost exclusively the dipoles situated just underneath them, on the cortex surface. Magnetometers on the other hand can record the magnetic field coming from more distant sources. In any case, the magnetic fields decay very rapidly when we move away from the source, as it is proportional to the squared distance between the source and the sensor. Therefore, there is a substantial loss in sensitivity for deep brain sources in MEG.

3.2.2 The sources of electro-magnetic brain signal

What do MEG and EEG record in the brain? MEG and EEG record, respectively, the magnetic and electric brain activity corresponding to the currents generated by the electric potentials of neurons. In the neuron, intracellular currents have two sources. Action potentials are responsible for rapid current flows along the axon. On the other hand, excitatory and inhibitory post-synaptic potentials that are produced along the dendrites and in the soma of neurons correspond to slower and more complex ionic currents in the extracellular space. It is possible to record both of these types of currents with local electrophysiological recordings and both types of current generate electromagnetic fields. However, they sum up in extracellular space resulting in the signal coming from a single cell being difficult to separate from activity generated by nearby cells. Therefore, when recording the electromagnetic field outside the scalp, the signal can only be the results of the activity of neuronal ensembles in which the neuronal currents are synchronous and therefore able to reach a level that is detectable by M/EEG sensors.

Importantly, the architecture of neuronal organization is thought to be a critical factor for M/EEG recordings. In particular, pyramidal cells in the layers of the cortex are arranged longitudinally, with cell bodies and axons oriented in a perpendicular way to the cortical surface. As nearby neurons are tightly interconnected, the currents of these cell-assemblies are thought to be at the origin of M/EEG recorded signals. Importantly, as action potentials are emitted very rapidly by neurons, it is unlikely that they would be synchronous enough to create the massive current flows recorded by M/EEG sensors. However, synchronous post-synaptic potentials (PSP), which correspond to a slower electric activity, might create long-lasting electromagnetic signals and therefore are considered as a more plausible source for M/EEG signals.

Importantly, currents also exist at the scale of the entire brain and correspond to two main types (Figure 3.2). Primary currents are those that are directly generated by neural assemblies reflecting synchronous PSP activity. These currents can be modelled as an equivalent current dipole in the corresponding brain regions and the electromagnetic field they create is recorded in a reliable manner by MEG, since magnetic activity is not distorted by other brain tissues, the skull or the skin. On the other hand, the secondary/volume currents correspond to larger currents that result from the interaction of all the primary currents and the head tissues. These volume currents, which occur at the scale of the head, are thought to be the primary source of the EEG signal (being responsible for its lower spatial resolution) and are also recorded by the MEG.

3.2.3 Reconstructing the source of M/EEG signal

One of the goals of M/EEG recordings is to try to reconstruct from the sensor signal the pattern of activity in the brain at the origin of the electromagnetic signal. This question corresponds to two distinct problems: the forward problem constitutes the understanding of what is measured with M/EEG devices. The inverse problem on the contrary is the procedure that consists of recovering the distribution of the neural generators that have produced the measurements.

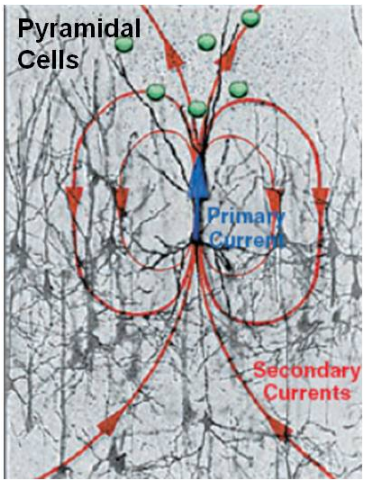


Figure 3.2: The two main types of currents in pyramidal cells

A state of the art manner of solving the forward problem is to make use of individual anatomical data from MRI scans. From the structural MRI of the subject, it is possible to extract the different structures of the head, such as the skull or the cortex surface, in order to obtain a detailed segmentation of the different elements composing the brain, in particular the white matter and the grey matter. It is then possible to locate where exactly in space are the sources of the currents that generated the M/EEG signals. Different head models need to be used in MEG and EEG. In particular, while MEG needs only the cortical surface and the location of the sensors to compute a forward model, EEG necessitates the modelling of the conductivity of the different tissues to estimate the forward model. Using a realistic modeling approach such as the boundary element method (BEM), it is possible to generate such a model from the precise MRI anatomical data.

Several methods exist to solve the inverse problem. In the distributed approach, the "source space" is constituted of dipoles placed all over the cortex, on the grey-matter surface, creating a 3D grid mesh of possible source points. Crucially, the orientation of the dipoles constituting the source of the signal can either be defined a priori, under the assumption that they are normal to the cortical surface, or on the contrary they can be left unconstrained. The architecture of the neuronal layers formed by the pyramidal neurons indicates that dipoles should be orientated perpendicularly to the cortical surface, therefore favouring the constrained option. However as brain segmentation from T1 MRI contrast might not always be perfectly accurate, it was proposed to adopt a loose constraint value, in order to take into account the imprecision of the segmentation of the cortical surface. Having defined the source space used, the minimum norm approach proposes a solution to solve the inverse problem. We can define the measurement as a linear transformation of the activity of the dipoles plus additional noise according to the following equation:

$$M = GD + N \tag{3.1}$$

where G represents the forward model, M is the measurement (EEG, MEG or both), D is the unknown dipoles activity and N is the noise.

Importantly, the M/EEG inverse problem with distributed source model is strongly ill posed as the number of source points is much greater than the number of sensors. The minimum-norm approach proposes to estimate D^* by solving the following optimization problem:

$$D^* = arg\ min \|M - GD\| \qquad subject\ to\ \|D\| \leq \eta \tag{3.2}$$

$\|M - GD\|$ represents the difference between the predicted measurement and the actual measurement and can be understood as a "reconstruction error" term. Therefore, this solution can be understood as trying to minimize a "reconstruction error" term, represented by $\|M - GD\|$ while imposing the solution $\|D\|$ to be smaller than a value η.

To overcome the loss of sensitivity for deep brain sources which occurs with MEG, it is possible to normalize the result by the sensitivity of the sensors, realizing a statistical test on the the source current according to the measure of the noise in a baseline condition. This method, dSPM, has been shown to reduce the loss of sensitivity for deeper sources. Furthermore, it has been proposed that a depth-weighting parameter could be applied, in order to give a higher weight to the signal originating from deeper brain regions. With these two parameters, we can show that even deeper brain sources can be estimated accurately. Indeed, when running a simulation in which a single dipole in the cingulate gyrus is active, as can be seen on Figure 3.3, the simultaneous use of dSPM and depth-weighting allows one to obtain a satisfying reconstruction of the activity in this deep region, unlike in a classic minimum norm estimate.

3.2.4 Why use simultaneous MEG/EEG recordings

Simultaneous MEG/EEG recordings constitute a powerful neuroimaging approach. First, M/EEG temporal resolution is of the order of the millisecond, allowing one to have a precise idea of the temporal dynamics of brain activity. In this regard, M/EEG offers a great advantage over fMRI, whose temporal precision is on the order of the second. However M/EEG spatial resolution remains poor compared to those of fMRI, as MEG source reconstruction does not permit one to go below the centimetre precision. Nevertheless, the advantage of these techniques is that cerebral activity is recorded instantaneously and simultaneously as a whole, contrarily to intracranial recording or fMRI.

Why use both MEG and EEG? EEG signals are mostly sensitive to volume currents flowing trough the head. In addition to lowering the spatial resolution of EEG, this imposes the necessity to generate a model of the conductivity of the different tissues of the head, which in many cases will be imprecise. On the contrary, MEG does not present such problem, resulting in greater spatial discrimination of neural sources. Considering these problems, it seems obvious that one should favour MEG and simply discard EEG recordings. As we have seen however, MEG and EEG are not sensitive to the same electromagnetic elements. In particular, EEG is much more sensitive to distant sources, which originate from deep brain

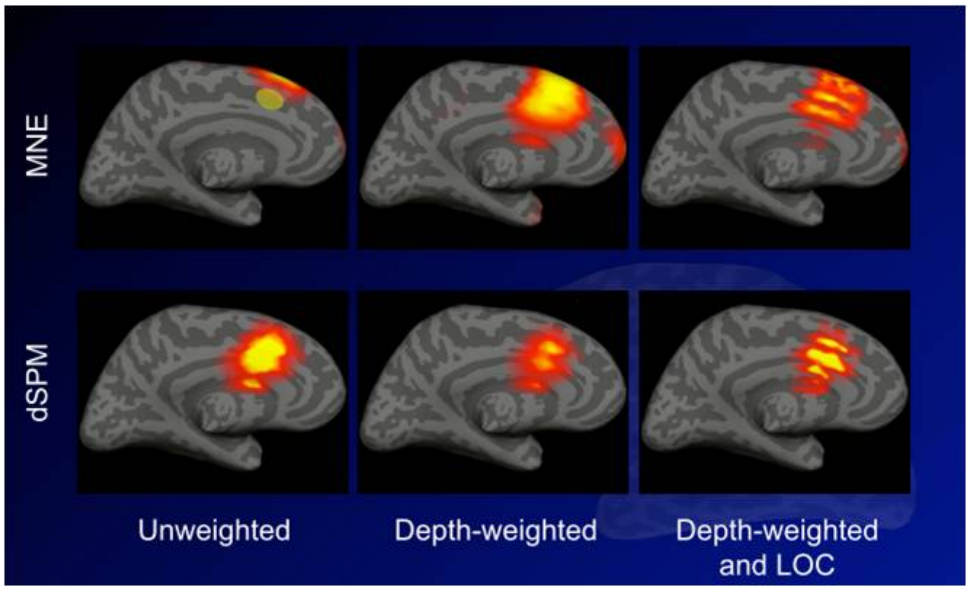


Figure 3.3: The effect of dSPM and depth weighting in the estimation of the source of a simulated dipole in ACC (from M. Hamalainen)

areas. Furthermore, MEG and EEG have a different degree of sensitivity to orientation of the dipoles in the brain (Figure 3.4). MEG sensors are not sensitive to dipoles oriented in a radial manner to the scalp and thus perpendicular to the sensors. Therefore, EEG signal can improve the sensitivity for these sources compared to MEG measurement alone. Additionally, ERP components in EEG have been very precisely documented, providing a reference when studying MEG signals. Finally, MEG and EEG can be easily combined to perform source reconstruction of both signals simultaneously, first computing a specific forward model for each type of signal and second normalizing the measures so that they are on the same scale. Therefore, the use of both techniques simultaneously allows one to obtain a very precise idea of the dynamics of brain activity in a given cognitive task

.

3.3 Decoding

In the present work, we adopted a decoding approach. Our goal was to identify with decoding tools how information used at the different stages of stimulus processing was modulated by consciousness. In the following section we discuss the theoretical background and the potential advantages and drawbacks of this method applied on M/EEG data.

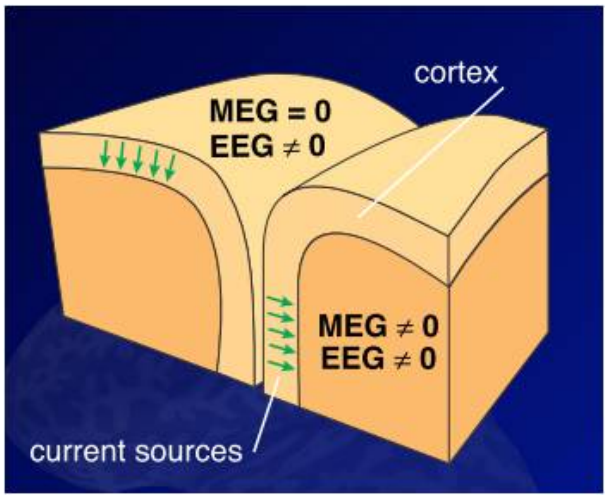


Figure 3.4: Relative sensitivity of EEG and MEG to dipole orientation in the head

3.3.1 Mulitvariate Pattern Analysis

The question of multivariate analysis of brain imaging data has been a recent focus of attention. In the univariate approach, tests are applied at one location of the brain or on a given cluster of sensors, allowing the investigation of the effect of the variable of interests on a specific element. This approach is highly relevant when a predefined brain region is the object of investigation, as is often the case in fMRI studies, or when the variable manipulated is predicted to affect a known evoked potential in EEG. However, this simple approach is not always appropriate.

In particular, an important question that needs to be addressed in neuroimaging research is: *"Is this specific information present in the brain and where?"*. This question focuses on the pattern of activity related to a cognitive operation, rather than the simpler question of the modulation by an experimental factor of a precise brain regions or ERP. Decoding provides a way to answer this question by transforming it into: *"Can we decode this specific information in brain activity and how?"*. In other words how much information do brain activity patterns carry and how do they relate to a precise mental state?

This approach has several advantages. First, it is blind to the experimental question addressed in the study and therefore should be less sensitive to the problem of double-dipping (Kriegeskorte et al., 2009). This confound occurs when we select the data on which we intend to perform our analysis using the same criteria as the hypothesis we would like to test. Indeed, it is easy to introduce this bias in the way we select a brain region or a set of sensors. The most common mistake consists in finding the region that is the most responsive for a condition A and then showing it indeed responds more for this condition A than for another condition B. Similar biases exists for ERPs analysis, for example when choosing a time-window for statistical analysis. When using decoding method, the analysis is partially blind to these confounds as the very question asked is whether the decoder can find the information "on its own", without the help of the experimenter.

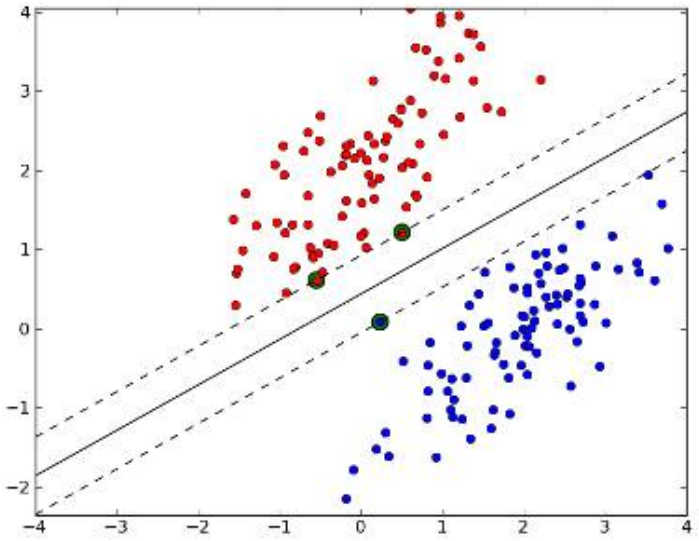


Figure 3.5: Linear classification with SVM in a two dimensional space. Black sold line constitutes the decision line that separates the two clouds of points.The margins are defined by the two dotted lines. Points inside the margins (green dots) constitute the support vectors (green dots).

The second advantage of decoding compared to classic methods is to overcome the complexity of the data, an issue that is particularly relevant for MEG. The general-linear model approach in fMRI developed by Friston et al. (1995) allows one to easily summarise the pattern of brain regions related to a specific contrast of conditions. However, EEG analysis requires one to deal with the additional time dimension. Furthermore, as MEG has an even larger number of sensors, dimensionality of the data increases drastically, making the use of classic univariate analysis difficult. While source reconstruction allows one to combine all the sensor-data, the question of how to incorporate the time dimension into the analysis remains an issue. Therefore, multivariate analysis offers a great potential for MEG data analysis.

3.3.2 Support Vector Machine

Several decoding approaches have been used in neuroimaging to decode brain activity. We focus here on Support Vector Machines (SVM). The SVM algorithm was developed in the early 90's (Boser et al., 1992) with the idea of finding an optimal linear classifier. To present linear classifiers and SVMs in particular, let us consider a two-dimensional dataset (Figure 3.5), in which each trial-by-trial data (each point of the figure) either belongs to a class A (points in red) or a class B (points in blue). In this framework, the idea of the linear classifier is simply to find a line that separates the two classes in this two-dimensional space.

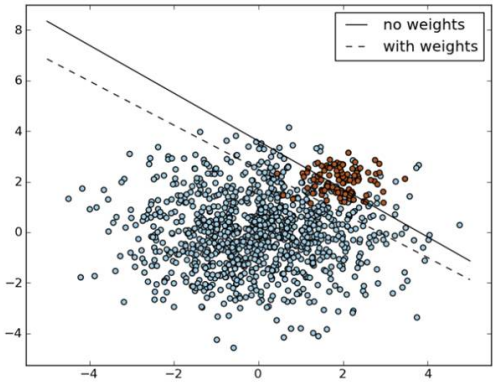


Figure 3.6: Problem of unbalanced dataset and sample-weight. When one class is more populated than the other, SVM tends to miss-classify more samples of the unpopulated class, failing to find an optimal decision line (solid-line). This effect can be corrected by applying sample weight, as shown by the obtained decision line (dashed line).

To find the linear classifier, the SVM algorithm uses only a small subset of the points that are the most informative. Counter-intuitively, the most informative points of the dataset are the ones that are the closest to the other points of the opposite class and the best line to separate the two classes is the one with the maximal distance to points of both classes (i.e. more separation between the classes). This distance from the line to the points of each class is called the margin and SVM can be defined as looking for the line separating the two classes with maximal margins. Indeed, a larger margin corresponds to a better generalization: if we add new points to the figure, the line that is the farthest from each cloud of points has a better chance of correctly classifying them.

The samples that lie on the margin are called support vectors, as they are the most difficult data points to classify and therefore define the location of the separating line. Interestingly, it was further proposed that instead of using the absolute maximum margin, it is possible to use "soft margins" (Cortes and Vapnik, 1995) which allow for the misclassification of some points, if the line cannot perfectly separate the data. This fitting process can be extended to a higher dimensional space. While the classification remains a linear process, the line is now called a hyperplane, which separates the data according to their different dimensions. In any case, the fitting process needs to be embedded in a cross-validation loop so that the classification of the data is meaningful and the data are not overfitted. The general recommendation in this regard is to separate the data into a training and a testing dataset which allows verification that the hyperplane can generalise its classification ability to unseen data. An optimal way to do so is to separate the data into stratified k-folds which are simply partitions of the samples that respect the proportion of each class.

Indeed, one major problem for any linear classifier occurs when the dataset is unbalanced and one class is more populated than the other. In this case, if we consider only the percentage of correctly classified samples, as chance level is not 50%, the classifier might ending up classifying one class very

well but missing many points of the unpopulated class. This can be overcome by using sample-weight, which weights maximally the points of the unpopulated class (Figure 3.6).

Taking this issue into account, SVM has proven to be a powerful tool for decoding. While it has been used more extensively in fMRI (McIntosh et al., 1996; Haynes and Rees, 2005; Norman et al., 2006), SVM method can also be applied to electro-physiological data. Importantly in this case, different features can be used to train the classifier. One option is to train a different classifier at each time point, using as a classification feature only the spatial dimension ($n_{channel}$ dimensions), i.e. the topographies of the M/EEG data. In this case, it is possible to obtain a classification score for each time-point and then reconstruct from these sample-by-sample data the entire time-course of classification accuracy for each trial, in order to study more precisely the dynamics of the related cognitive process. Of course in this case, it is important to keep in mind that a distinct classifier is computed for each time-sample. Therefore, any pattern that is jittered across time might fail to be classified, not because the information is absent but simply because it is not present at the same instant in all trials, and thus impossible to be picked up by the classifiers.

Another possibility is to train the classifier on a larger time-window, giving both time and space as decoding features. In this case, the decoder learns to decode in a high dimensional space, with a total of $n_{time-point} * n_{channel}$ dimensions. Importantly, the output of the classifier is then simply the classification value for each trial which does not allow the assessment of the dynamics of the decoded process. In particular, this approach will not permit the determination of which period was used by the decoder. However, it can sometimes improve the decoding accuracy, for example when the information is slightly diluted in time.

3.3.3 Evaluating classification score

When a given classifier has been computed, we usually analyse the classification values i.e. to which class a given trial has been assigned. Alternatively, it has been proposed that we can compute the probability for the trial to belong to one category or the other. This can be achieved by fitting a sigmoid logit function onto the output of the classifier (Platt, 1999). It is possible to obtain for each trial the decision variable which gives an estimate of how far this trial is from the classification boundary. Each trial can then take a value 0 or 1 according to on which side of the boundary it falls. It is then possible to fit the logit function on this graph, linking the decision variable to the classifier output. Of course, such a step has to be done in the cross-validation loop in order to avoid over-fitting.

In this way, it is then possible to obtain a classification measure for each trial that can vary continuously between 0 and 1 instead of a binary measure. This method presents several advantages. First, it allows one to have a clearer idea of the classification performance and in particular of the variations across trials. Second, it allows one to use more appropriate statistical measures to compute the significance of the classification. In particular, having obtained the distribution of probability across all trials allows one to perform statistical analysis such as computing the ROC curve, which is known to be a

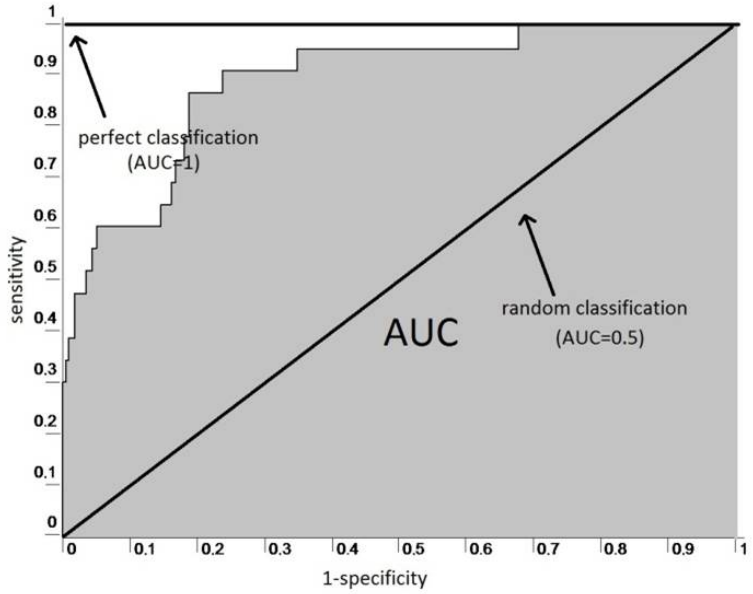


Figure 3.7: A measure of classification sensitivity: AUC and ROC curve. ROC curves plot the sensitivity of the classification score versus its specificity and can be estimated from the probability distributions of the output of the classifier. The area under the ROC curve (AUC) provides a measure between 0 and 1 of the classification score.

powerful measure of classification sensitivity. As the probability distributions for hits and false alarm are known in this case, the ROC curve can be computed by plotting the cumulative distribution function (area under the probability distribution) of the detection probability in the y-axis versus the cumulative distribution function of the false alarm probability in x-axis (Figure 3.7). From the ROC curve, it is possible to compute the Area-Under Curve (AUC) value that represents how much the sensitivity differs from chance. A diagonal ROC curve, which coincides with an AUC of 0.5, corresponds to a situation where the numbers of hits and false alarms are equal, showing a chance level classification score. On the contrary, an AUC of 1, which corresponds to a ROC curve on the left upper bound of the diagonal, indicates a perfect positive prediction with no false positives and a perfect decoding score (Figure 3.6). Importantly, unlike average accuracy, AUC analysis provides an unbiased measure of decoding accuracy, robust to imbalanced problems and independent of the statistical distribution of the classes. It also allows one to compute significance of the classification sensitivity, this statistic also being robust to problems of unbalanced data.

3.3.4 The benefit and confounds of decoding

Decoding is a powerful technique for many reasons. In addition to its high statistical power compared to classic statistical methods, it allows one to address questions that are particularly relevant for cognitive neuroscience. An important aspect of this resides in the training/testing approach that is used in supervised learning. We try to list some of these features below.

1. Decoding allows the determination of whether a given piece of information is present or not in brain activity.

2. Decoding provides information on which region or at which point in time the information of interest is present.

3. Decoding can be used to determine how well one cognitive process generalises to other conditions.

The first point is the most intuitive. Multivariate decoding provides a simple approach to exploring data and determining if a specific information content is present in brain activity. We have seen that it allow two difficulties to be overcome: first it avoids "double-dipping" strategies in which scientists are tempted to use circular analysis to find the effect of interest (Kriegeskorte et al., 2009). This type of analysis whereby one (involuntarily) inserts distortions in the results by selecting a subset of the data for analysis can occur very easily, in particular when an abundant literature reports the same effect. Such an approach can lead to the systematic negligence of significant results and the selection of expected results to the detriment of the objective result patterns. A second point tightly linked is that decoding provides a way to overcome the problem of very large amounts of data. This is particularly relevant for M/EEG which associates a great number of sensors (306 MEG sensors and 60 EEG sensors) to an excellent time-resolution (1kH). Considering these datasets, it is difficult to obtain an overall vision of the data and the validity of one given approach compared to another to determine the existence of a precise effect. In this aspect, decoding of binary conditions allows the determination of whether a significant difference between two conditions exists or not.

Nonetheless, conclusions drawn using decoding techniques should be considered with caution. For instance, it is difficult to conclude on chance-level decoding performance. Beyond the fact that null results ought to be considered as a lack of evidence rather than a strict proof of absence of effect, classification results are tightly limited by the decoding tools that are used, and poor classification scores do not prove an absence of information concerning the related cognitive process. Therefore, the inability to learn to classify a specific pattern cannot be regarded as a strict proof of the absence of information concerning the related cognitive process. On the other hand, very high decoding scores to classify two conditions from one another need to be regarded critically. In particular, as decoding techniques are blind to the dimensions used to classify the data, any difference between two conditions can be picked-up by the decoder, whether or not it is relevant for the cognitive question addressed. Such criticism have been carefully presented in several articles (Todd et al., 2013; Lemm et al., 2011): for example if the two conditions that are classified correspond to two distinct blocks, it is enough for the decoder to pick-up information concerning the block (such as noise level or baseline shift) to classify the two conditions, without decoding anything related to the difference in brain activity.

The second point regarding the power of decoding is that the decoder provides information on which features might be used to predict to which class each trial belongs. This point is one of the most problematic. As we have seen, any relevant information can be picked-up by multivariate pattern decoding,

making it a useful tool with which to detect subtle differences between conditions. However, no constraints exist on which information is going to be used: in particular it might be possible to decode the difference between two conditions in a brain area that is not directly relevant for the cognitive function studied. Therefore, the sensors or the brain region used by the decoder to classify two classes should not be considered as reflecting the underlying maximal region of activity for the studied cognitive process. Another point is that multivariate pattern classifier may discard information that is redundant in patterns of brain activity, leading sometimes to the belief that patterns of relevant information consist of smaller regions or clusters of sensors than the ones actually carrying the information. A good way of understanding this problem is to consider when one is trying to decode a response hand in a binary motor task. As only two responses are possible in this task, it is enough for a decoder to look at one of the two motor cortices to determine which hand has been used, each lateralized motor cortex behaving as an on/off signal of activity for each hand. Looking at which region is used by the decoder in this case could lead to the belief that only one region of motor cortex is relevant while in fact both left and right regions are active in the task.

A last important point that should be noted is that, as a result of the training and testing approach, decoding allows the study of how one classifier can generalize its classification ability to a new problem. This is particularly interesting as it allows one to determine how decoding of a given condition can be generalized to another, giving an estimate of the common information shared by the two conditions. Again, results of generalisation should be considered with caution. In particular, the inability to generalise from one condition to another cannot be considered as strict proof of the absence of common information between the two. However, the possible generalisation from one condition to another provides an index of the degree of shared information between them.

3.4 Followed plan

In the experimental part of this thesis, three articles are presented that investigate the link between consciousness and metacognition.

In the first study, we asked whether mechanisms of error-detection can be triggered non-consciously. Metacognition has been linked to consciousness, following the hypothesis that processes that can be introspected should be conscious (Kolb and Braun, 1995; Rounis et al., 2010; Lau and Passingham, 2006; Persaud et al., 2007). However, evidence exist of brain activity and behaviour related to complex cognitive control functions tightly linked to metacognition that occur outside of consciousness (Cohen et al., 2009; van Gaal et al., 2008; van Gaal et al., 2009; Lau and Passingham, 2007; Pessiglione et al., 2008; Pessiglione et al., 2007). Therefore, the question of whether metacognitive information can be extracted in non-conscious conditions should be tested. In particular, our goal was to investigate two aspects. First, very few studies have investigated in an objective manner performance in error detection tasks in non-conscious conditions. While research of subliminal processing has developed forced-choice tasks and objective measures of detection sensitivity to assess the level of information available on subliminal

stimuli, only subjective confidence ratings have been asked so far to assess metacognitive knowledge about the accuracy of decisions. We proposed in this study to assess meta-performance in a forced-choice task on accuracy, in conscious and non-conscious conditions. Second, no consensus exist on whether brain signals related to error detection are present in non-conscious conditions (Woodman, 2010; Pavone et al., 2009; Nieuwenhuis et al., 2001). We proposed to test how one of this markers, the ERN, was modulated both by subjective visibility reports and objective variation in masking strength, in order to obtain a clearer idea of the impact of conscious processing on known performance monitoring processes. In a first series of experiments, we showed that the ERN was absent in subliminal conditions in which stimuli were presented too briefly for subjects to detect their occurrence. Surprisingly however, we found that subjects were still able to report their performance better than chance(Charles et al., 2013) in non-conscious trials in which they denied seeing the stimulus, indicating an interesting computational difference between confidence judgments and all-or-none error-detection in regard to consciousness.

Following this surprising results, we tried to address the question of the nature of the difference between conscious and non-conscious error monitoring processes. In particular, we hypothesized that the performance monitoring mechanisms in conscious and non-conscious conditions might extract information on the accuracy of the decision in two distinct ways: while non-conscious performance monitoring might correspond to a statistical assessment of confidence in the response, error-detection in conscious conditions might be based on a categorical judgement resulting from the comparison of intended and executed actions. According to this view, conscious trials would distinguish from non-conscious trials by the emergence of a clear intention signal, representing the correct required action, that would still be present when committing an error even though it arrives too late to influence directly the motor output. To test this hypothesis, we used decoding methods of SVM linear classifiers described above to contrast patterns of brain activity associated with particular cognitive processes and behaviours linked to the dynamics of action selection and performance monitoring. In particular, we isolated brain activity information related to the computation of the correct/intended response, independently of the actual motor response produced by the subject, and determined how it was modulated by consciousness (Charles et al., 2013). We found that information related to the intended response could be decoded in brain activity only in conscious trials, as predicted by our model. Furthermore, we found that accuracy of the motor decision could be decoded at a time and with an accuracy that depended on the decodability of the required and the executed actions. These findings led us to propose an alternative model of error detection that would rely on the comparison of two streams of information: non-conscious computation of the motor response and conscious computation of the required response.

In a third study, we further tested whether conscious and non-conscious metacognitive processes were truly distinct. To do so, we replicated our initial protocol in a population of schizophrenic patients. Indeed, schizophrenic patients are known to present specific deficits in conscious conditions while their non-conscious processes seem to remain unimpaired. Interestingly, the ERN has been shown to be reduced in schizophrenia, as predicted by our model of the ERN depending on conscious access. Our prediction was the following: if conscious and non-conscious performance-monitoring processes truly

dissociate, then conscious error detection should be impaired in schizophrenia while non-conscious confidence judgement should be preserved. First, we replicated our previous results showing that metacognitive performance could remain above-chance in non- conscious conditions while the ERN was present only in conscious conditions. More importantly however, we found that schizophrenic patients presented similar metacognitive performance in subliminal conditions as in control subjects although conscious error detection processes were altered (Charles and Dehaene, 2013). These results show that schizophrenia is associated with a deficit in conscious access as previously found and constitutes a proof case demonstrating that indeed metacognitive processes deployed consciously and non-consciously are computationally distinct.

Part II

Experimental contributions

CHAPTER 4

Article 1 : Distinct brain mechanisms for conscious and subliminal error detection

---

4.1 Introduction to the article

4.1.1 Context and goal of the study

While it has been proposed that consciousness and metacognition are tightly linked (Persaud et al., 2007), the question of whether they dissociate in some cases remains to be tested. We have seen that some evidence can be accumulated on a masked stimulus even when it is not consciously detected (Del Cul et al., 2007). Furthermore, some higher-order cognitive functions that would be intuitively linked to conscious experience can be triggered non-consciously (van Gaal et al., 2008; van Gaal et al., 2009; Pessiglione et al., 2007). Therefore, the question of whether metacognitive information can be extracted in non-conscious condition remains crucial. Importantly, it has been proposed that confidence (Rounis et al., 2010) as well as the ability to wager on the accuracy of responses (Persaud et al., 2007) constitutes an assessment of subjective awareness, with the underlying assumption that this information is definitively unreachable by conscious access. However, in the same manner that subjective reports of visibility have been criticized for their lack of insight on subjects' response bias and thus led to the development of more objective measures, the question of whether in a forced-choice task, metacognitive sensitivity can be better than chance remains to be tested.

Some evidence exists that error-detection can operate non-consciously. In particular, it was shown that when subjects failed to detect their own incorrect movement and therefore missed making an error, an ERN could still be observed (Nieuwenhuis et al., 2001; Endrass et al., 2007). Therefore, it can be established that the ERN can be triggered for unaware actions. However, can an ERN be evoked when performing a task on subliminal stimuli? This question remains unsettled: while some authors found that the ERN is indeed absent when responding to a masked stimulus (Woodman, 2010), others have found that a weak negativity is still present (Pavone et al., 2009) but only marginally different from the negativity in correct trials. Furthermore, several studies found that the ERN varies with subjective confidence in the correct response (Pailing and Segalowitz, 2004a) suggesting that conscious perception of the stimulus should influence the presence of the ERN.

4.1.2 Experiment

To address this question, we used a masking paradigm similar to that of Del Cul et al. (2007) (see part I) in order to obtain different condition of visibility. Importantly we asked the subjects to perform three tasks on each trial: an objective task on which we applied time-pressure, the goal being to study error processing on this specific task, time-pressure causing the subjects to commit a lot of errors; a subjective visibility task (seeI) and an error-detection task in which participant had to say in a binary manner if they thought they made an error or not. In order to ensure that our results were not simply caused by the time-pressure imposed in the objective task, we conducted two experiments in which time-pressure was varied being either very strong or more relaxed.

Our goal was to perform two main analyses:

1. Test whether the ERN can be observed in trials where the subject reported that he or she did not see the stimulus. In particular, considering the exact same stimulation condition, when masking strength is kept constant, how does subjective report influence the amplitude of the ERN? This analysis can be achieved by separating the trials by both visibility reports and SOA condition.

2. Test how do subjects perform in detecting errors in the trials in which they report not seeing the stimulus. In particular, for trials in which, as in Del Cul et al. (2007) partial accumulation of evidence can be achieved on the subliminal stimulus, can subjects also predict their performance?

4.1.3 Summary of the results

Interestingly our study revealed a striking dissociation:

- On the one hand, the ERN was affected in an all-or-none fashion by visibility: it was present only when subjects reported consciously perceiving the stimulus even when the stimulus was kept constant, and the trials differ only in subjective visibility versus invisibility.

- On the other hand, even in the absence of ERN, we observed that subjects remained better than chance in evaluating their performance in non-conscious conditions, demonstrating a form of non-conscious meta-cognition

Taken together, our findings identified two distinct processes at work in conscious and non-conscious conditions: all-or-none error detection, indexed by the ERN is present only in conscious conditions, but confidence in one's response can still be computed under non-conscious conditions. Our results therefore strongly supports the view that some high-level processes are "all-or-none" and activate only under conscious conditions but that at the same time non-conscious metacognitive information exists, while relying on distinct brain processes.

4.2 Article

Charles, L., van Opstal, F., Marti, S. & Dehaene, S. 2013 Distinct brain mechanisms for conscious versus subliminal error detection. NeuroImage 73, 80-94.

NeuroImage 73 (2013) 80–94

Contents lists available at SciVerse ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Distinct brain mechanisms for conscious versus subliminal error detection

Lucie Charles [a,b,c,*], Filip Van Opstal [a,b,c,d], Sébastien Marti [a,b,c], Stanislas Dehaene [a,b,c,e]

[a] INSERM, U992, Cognitive Neuroimaging Unit, CEA/SAC/DSV/DRM/NeuroSpin, Bât 145, Point Courrier 156 F-91191 Gif/Yvette, France
[b] CEA, DSV/I2BM, NeuroSpin Center, Bât 145, Point Courrier 156 F-91191 Gif/Yvette, France
[c] Univ Paris-Sud, Cognitive Neuroimaging Unit, Bât. 300-91405 Orsay cedex
[d] Ghent University, Henri Dunantlaan 2, B-9000 Ghent, Belgium
[e] Collège de France, 11, place Marcelin Berthelot, 75231 Paris Cedex 05, France

ARTICLE INFO

Article history:
Accepted 26 January 2013
Available online 4 February 2013

Keywords:
Error-related negativity
Consciousness
MEG
EEG

ABSTRACT

Metacognition, the ability to monitor one's own cognitive processes, is frequently assumed to be univocally associated with conscious processing. However, some monitoring processes, such as those associated with the evaluation of one's own performance, may conceivably be sufficiently automatized to be deployed non-consciously. Here, we used simultaneous electro- and magneto-encephalography (EEG/MEG) to investigate how error detection is modulated by perceptual awareness of a masked target digit. The Error-Related Negativity (ERN), an EEG component occurring ~100 ms after an erroneous response, was exclusively observed on conscious trials: regardless of masking strength, the amplitude of the ERN showed a step-like increase when the stimulus became visible. Nevertheless, even in the absence of an ERN, participants still managed to detect their errors at above-chance levels under subliminal conditions. Error detection on conscious trials originated from the posterior cingulate cortex, while a small response to non-conscious errors was seen in dorsal anterior cingulate. We propose the existence of two distinct brain mechanisms for metacognitive judgements: a conscious all-or-none process of single-trial response evaluation, and a non-conscious statistical assessment of confidence.

Introduction

What are the limits of non-conscious processing? In the past twenty years, evidence has accrued in favor of deep processing of subliminal stimuli (i.e., stimuli presented below the threshold of subjective visibility). Not only can early visual processing be preserved under masking conditions (Del Cul et al., 2007; Melloni et al., 2007), but subliminal primes can modulate visual (Dehaene et al., 2001), semantic (Van den Bussche et al., 2009) and motor stages (Dehaene et al., 1998; for a review, see Kouider and Dehaene, 2007). Even executive processes, once considered the hallmark of the conscious mind, can be partially influenced by non-conscious signals related to motivation (Pessiglione et al., 2007), task switching (Lau and Passingham, 2007) and inhibitory processes (Van Gaal et al., 2008). These findings raise the issue of whether subliminal stimuli could affect *any* cognitive process, or whether certain processes depend on an all-or-none conscious ignition (Del Cul et al., 2007).

Here, we investigate meta-cognition — the ability to reflect on oneself and on one's own cognitive processes. Intuitively, introspective reflection is virtually indistinguishable from conscious processing: it is hard to envisage introspection without consciousness. This intuition has served as a basis for the frequent identification of consciousness with self-oriented, metacognitive or "second-order" cognition: any information that can enter into a higher-order thought process would be conscious by definition (Kunimoto et al., 2001; Lau and Rosenthal, 2011; Persaud et al., 2007). However, this conclusion may also be disputed. Some metacognitive monitoring processes, such as those associated with the evaluation of one's performance (Logan and Crump, 2010) or the subsequent correction of one's errors (Endrass et al., 2007; Nieuwenhuis et al., 2001; Wessel et al., 2011) are conceivably sufficiently simple and automatized to be deployed non-consciously. Thus, whether metacognitive processing implies conscious processing can and should be tested empirically.

To investigate how performance monitoring relates to conscious perception, the present experiments concentrate on the error-related negativity (ERN), a key marker of error processing. The ERN is an event-related potential that peaks on fronto-central electrodes 50 to 100 ms after making an erroneous response; it is easily observed in EEG recordings (Dehaene et al., 1994; Falkenstein et al., 2000; Gehring et al., 1993), and a similar, though harder to detect MEG component has been reported (Keil et al., 2010; Miltner et al., 2003). The ERN is assumed to originate in the cingulate cortex (Agam et al., 2011; Debener

Abbreviations: ERN, Error-Related Negativity; ERP, event-related potential; ERF, event-related field; SDT, Signal Detection Theory; SOA, stimulus onset asynchrony; MEEG, simultaneous magneto- and electroencephalography.

* Corresponding author at: INSERM-CEA Cognitive Neuroimaging unit CEA/SAC/DSV/DRM/NeuroSpin Bât 145, Point Courrier 156 F-91191 Gif/Yvette, France. Fax: +33 1 69 08 79 73.
E-mail address: lucie.charles.ens@googlemail.com (L. Charles).

1053-8119/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neuroimage.2013.01.054

et al., 2005) and its role in cognitive control has been related to error detection (Gehring and Fencsik, 2001; Nieuwenhuis et al., 2001), reinforcement learning (Holroyd and Coles, 2002) and conflict processing (Botvinick et al., 2001; Veen and Carter, 2002).

The debated issue that we address here is whether the ERN indexes a process which is automatic enough to be deployed unconsciously. In relating this issue to the existing literature, it is crucial to keep in mind that an error can fail to be consciously detected for several reasons. A distinction must be made between errors that remain unnoticed (1) because the erroneous action itself is not detected (for instance because it consists in a fast key press or eye-movement (Endrass et al., 2007; Nieuwenhuis et al., 2007; Logan and Crump, 2010; Hughes and Yeung, 2011)), (2) because the subject cannot determine which response is the correct one (e.g. when responding to a visible but confusing stimulus or instruction), or (3) because the subject is completely unaware of the stimulus and therefore of the correct response (e.g. when responding to a stimulus made invisible by masking).

Initially, the relationship between consciousness and the ERN was explored in the context of case (1), i.e. unaware actions (Nieuwenhuis et al., 2001). It suggested that the ERN may remain present even when participants are unaware of having made a partially erroneous eye-movement (Endrass et al., 2007; Nieuwenhuis et al., 2001; but see Wessel et al., 2011). In these studies, crucially, subjects performed a difficult antisaccade task and were sometimes unaware of their erroneous glances in the pro-saccade direction. These results were further extended to case (2) (i.e., confusion about which response is the correct one), in paradigms where undetected errors were induced by conflicting stimuli evoking two contradictory responses (Dhar et al., 2011; Hughes and Yeung, 2011; O'Connell et al., 2007 but see Maier et al., 2008; Steinhauser and Yeung, 2010). These studies have typically used the Eriksen flanker task, in which the presence of multiple conflicting letters may purposely confuse the participant as to the nature of the correct response.

Here, however, we aimed at testing the third case, i.e. whether an ERN can be elicited by an unseen masked stimulus. Our main motivation was to extend the existing literature on the depth of subliminal processing of masked words and digits (Kouider and Dehaene, 2007). In masking experiments, it is well known that participants may deny seeing the stimuli, yet still perform above chance level in a broad range of categorization task, such as deciding whether a digit is larger or smaller than 5 (Dehaene et al., 1998; Del Cul et al., 2007). As an extreme case, in blindsight, a patient may deny any conscious experience, while remaining able to perform way above chance in simple tasks on stimuli presented in their blind hemi-field (Kentridge and Heywood, 1999; Weiskrantz, 1996).

The specific question for the present research is whether, in subliminal conditions induced by masking, the error detection system may also be triggered non-consciously. We evaluate this question both by monitoring the presence of the ERN, as well as by asking the participants for a second-order behavioral response. On each trial, the participant first makes a forced-choice number comparison, and is then asked to decide whether he made an error or not. The finding of either an unconscious ERN, or of an above-chance second-order metacognitive performance on subliminal trials, would expand the range of unconscious operations. Corroborating recent evidence that even executive processes of task switching and response inhibition may be partially initiated non-consciously (Lau and Passingham, 2007; van Gaal et al., 2008), it would indicate that an unseen masked stimulus is capable of progressing through a hierarchy of successive processing stages, all the way up to a level of metacognitive monitoring. A negative answer, on the other hand, would support the view that there are sharp limits to unconscious processing, and that some cognitive operations only proceed once the stimulus has crossed an all-or-none threshold for conscious access (Aly and Yonelinas, 2012; Dehaene and Changeux, 2011; Province and Rouder, 2012; Sergent and Dehaene, 2004a).

Only two studies (Pavone et al., 2009; Woodman, 2010) investigated the existence of an ERN on subliminal trials, yet they obtained contradictory results: Woodman (2010) found that the ERN was absent for masked stimuli, while Pavone et al. (2009) found that it could still be detected. Crucially, in order to contrast conscious versus non-conscious processing, both studies manipulated parameters of contrast or duration. Such sensory manipulations *per se* can have a large impact on the amount of information available on subliminal trials compared to conscious trials. Their findings may therefore result in a large part from this objective change in stimulus strength. One of our aims was therefore to determine if changes in subjective perception alone, in the presence of a constant stimulus, would modulate the ERN and metacognitive performance. To this end, we measured error responses to visual stimuli of variable masking strength, ranging from fully visible to fully invisible (Fig. 1). Such design allowed us to determine how subjective perception of a stimulus, by itself, affects performance-monitoring processes, as assessed by behavioral and error-related MEEG brain measures.

In two masking experiments, participants performed a number comparison task on a masked digit, while perceptual evidence was systematically manipulated by varying the target-mask Stimulus Onset Asynchrony (SOA; Del Cul et al., 2007). To maximize the number of errors, a strong pressure to respond fast was imposed in experiment 1. The main results were replicated in a second experiment in which this pressure was reduced. Crucially, subjective perception was assessed on a trial by trial basis by asking participants to report their visibility of the target (*Seen*/*Unseen*) as well as their perceived performance (*Error*/*Correct*) in the number comparison task. Given that subjective reports vary spontaneously across trials, this approach allowed us to study how the ERN and error-detection performance were modulated by subjective perception of the stimulus (subliminal/subjectively *unseen* trials versus conscious/*seen* trials), independently of the objective variation in masking strength.

Materials & methods

Participants

In the first experiment, seventeen volunteers were tested (5 women and 12 men; mean age 23.8 years). Because our experimental conditions were partially determined by subjective reports, four participants were discarded for having insufficient numbers of trials in some of the conditions. Specifically, we removed participants with false-alarm rate superior to 10% in the mask-only condition, or with less than 15% of *seen* trials in the 50 ms SOA condition. In the second experiment, sixteen participants were tested (6 women and 10 men; mean age 23.2 years). Two had to be discarded due to technical problems during MEG recording. One participant was discarded using the same behavioral criteria as in the first experiment. In the end, each experiment comprised data from 13 participants. All participants had normal or corrected-to-normal vision.

Design & procedure

A masking paradigm similar to Del Cul et al. (2007) was used in this experiment. The target-stimuli (the digits 1, 4, 6, or 9) were presented on a white background screen using E-Prime software. The trial started with a small increase in the size of the fixation cross (100 ms duration) signalling the beginning of the trial. Then the target stimulus appeared for 16 ms at one of two positions (top or bottom, 2.29° from fixation), with a 50% probability. After a variable delay, a mask appeared at the target location for 250 ms. The mask was composed of four letters (two E's and two M's, see Fig. 1) tightly surrounding the target stimulus without superimposing or touching it. The stimulus-onset asynchrony (SOA) between the onset of the target and the onset of the mask was varied across trials.

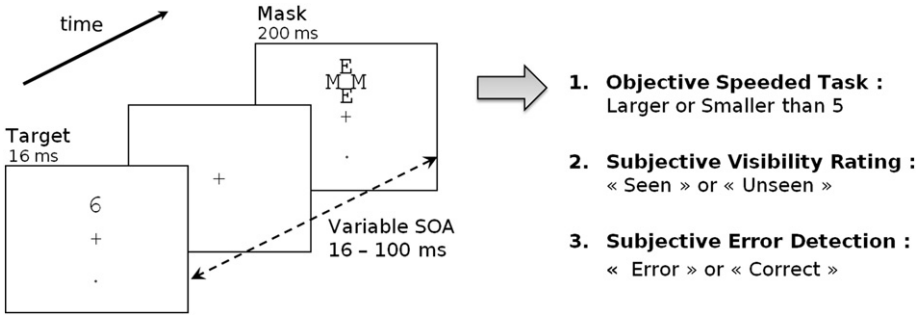


Fig. 1. Experimental design: On each trial, a number was presented for 16 ms at one of two possible locations (top or bottom). It was followed by a mask composed of a fixed array of letters centered on the target location. The delay between target onset and mask onset (SOA) varied randomly across trials (16, 33, 50, 66 or 100 ms). In one sixth of the trials, the mask was presented alone (mask only condition). Participants first performed an objective forced-choice number comparison task where they decided whether the number was smaller or larger than 5. In experiment 1, the response had to be made in less than 550 ms, otherwise a negative sound was emitted. In experiment 2, participants were simply instructed to respond as fast as they could while maintaining accuracy. Then, on each trial, participants performed two subjective tasks. First they evaluated the subjective visibility of the target by choosing between the words "*Seen*" and "*Unseen*", displayed randomly either left or right of fixation. Second, they evaluated their own performance in the primary number comparison task by choosing between the words "Correct" and "Error", again displayed randomly either left or right.

Five SOAs were randomly intermixed: 16, 33, 50, 66 and 100 ms. The foreperiod duration was manipulated so that the mask always appeared 800 ms after the signal of the beginning of the trial. In one sixth of the trials, the target number was replaced by a blank screen with the same duration of 16 ms (mask-only condition), allowing us to study visibility ratings when no target was presented.

Participants primarily had to perform a forced-choice task of comparing the target number to the number 5. Responses were collected within 1000 ms (experiment 1) or 2000 ms (experiment 2) after target onset with two buttons using the index of each hand (left button press = smaller-than-5; right button-press = larger-than-5 response). To induce errors, participants were instructed to respond as fast as they could just after the appearance of the target. In experiment 1, time pressure was increased by presenting an unpleasant sound (mean pitch: 136.2 Hz, 215 ms duration) 1000 ms after target presentation whenever response time exceeded 550 ms. In experiment 2, no further time pressure was imposed.

At the end of each trial, after another delay of 500 ms, participants were requested to provide two subjective answers with no time–pressure. The first answer was related to the subjective visibility of the target number. In this visibility task, participants had to indicate if they saw a target number or not. The second answer concerned the participants' knowledge of their performance. Here, they had to indicate whether they thought they had made an error or not in the number comparison task (performance evaluation task). Instructions were clearly stated to ensure that participants understood that the performance evaluation task was directed to the number comparison task and not the visibility judgment. Furthermore, participants were informed that, even when they had not seen the stimulus and thought that they responded randomly, they still had a 50% chance of having made a correct response. Therefore, they were told to hazard a guess on their performance, even when they did not see the stimulus. For both subjective responses, words corresponding to the two responses (*seen*/*unseen* and *error*/*correct*) were displayed on the screen and participants had to use the corresponding-side buttons to answer. The words were presented at randomized left and right locations (2.3° from fixation) to ensure that participants didn't use automatized button-press strategy.

The experiment was divided in blocks of 96 trials. Each block contained 16 trials for every SOA condition, with each digit presented at the two possible target locations (top/bottom). Participants performed 6 or 7 blocks during EEG/MEG recording. For Experiment 1, in order to achieve fast responses, participants were given a training session before the actual recording. They first received 5 min of training where the target stimulus was not masked. Next, participants performed 3 pre-recording blocks of the actual experiment in order to check that overall performance was suitable for MEG/EEG recording. In Experiment 2, where fast responding was not required, only ten trials of the experiment were given as training before starting the actual recording.

Simultaneous EEG and MEG recordings

Simultaneous recording of MEG and EEG data was performed. The MEG system (the Elekta-Neuromag) comprised 306 sensors: 102 Magnetometers and 204 orthogonal planar gradiometers (pairs of sensors measuring the longitudinal and latitudinal derivatives of the magnetic field). The EEG system consisted of a cap of 60 electrodes with reference on the nose and ground on the clavicle bone. Six additional electrodes were used to record electrocardiographic (ECG) and electro-oculographic (vertical and horizontal EOG) signals.

A 3-dimensional Fastrak digitizer (Polhemus, USA) was used to digitize the position of three fiducial head landmarks (Nasion and Pre-auricular points) and four coils used as indicators of head position in the MEG helmet, for further alignment with MRI data. Sampling rate was set at 1000 Hz with a hardware band-pass filter from 0.1 to 330 Hz.

SDT analysis

To obtain an unbiased measure of visibility and performance, we used Signal Detection Theory (SDT) to compute $d' = z(\text{HIT}) - z(\text{FA})$ for the target-detection task (*detection-d′*, where HIT = proportion of trials with target present and response *seen*, and FA = proportion of trials with target absent and response *seen*) and the number comparison task (where HIT = proportion of trials with target smaller than 5 and a left response, and FA = proportion of trials with target larger than 5 and a left response).

The *meta-d′* measure was computed according to Maniscalco and Lau (2012). Briefly, classic SDT can be extended to predict what should be the theoretical performance in meta-cognitive judgements where one must evaluate one's own primary performance, such as confidence ratings or error detection. The theory assumes that both primary and meta-cognitive judgements have access to the same stimulus sample on the same continuum. First-order judgments are performed by setting a first criterion in the middle of the continuum. Meta-cognitive judgements are performed by setting two additional criteria surrounding the first-order one, and responding "error" if the sample falls between these two criteria, or "correct" if the sample falls beyond them (i.e. a sample distant enough from the first-order criterion signals high confidence in the primary response). From this ideal-observer theory, precise mathematical relations linking performance and

meta-performance can be deduced (Galvin et al., 2003) and it is possible to compute a second-order measure of meta-performance by classifying meta-cognitive responses as second-order hits and false alarm. However, the traditional measure of d' does not directly apply to a second-order task because it is not unbiased (second-order d' systematically depends on the first-order criterion) and the assumption of normality of the distributions is violated. In order to obtain a valid measure of meta-performance, unbiased and comparable to the first-order d', Maniscalco et al. (http://www.columbia.edu/~bsm2105/type2sdt/) proposed an alternative solution, *meta-d′*. Their proposal consists in bringing both first and second-order performance to the same scale, by determining what should have been the d' in the first-order task given the observed second-order (meta) performance, under the assumption that the subject used exactly the same information in both cases. Since *meta-d′* is expressed in the same scale as d', the two can be compared directly. When *meta-d′* $< d'$, it means that the subject did worse in the performance evaluation task than expected according to his actual d' value. On the opposite, if the *meta-d′* $> d'$, it means that more information was available for subjective performance evaluation than for the primary objective decision.

Meta-d′ was estimated by fitting the parameters of a type-I SDT model so that the predicted type-II hits and false-alarm rates were fitted to the actual type-II data. Therefore, *meta-d′* corresponds to the d' that maximizes the likelihood of the observed type performance, assuming the same bias of response as the one observed in the data.

MEG/EEG data analysis

MEG data were first processed with MaxFilter™ software using the Signal Space Separation algorithm. Bad MEG channels were detected automatically and manually, and interpolated. Head position information recorded at the beginning of each block was used to realign head position across runs and transform the signal to a standard head position framework.

To remove the remaining noise, Principal Component Analysis (PCA) was used. Artifacts were detected on the electro-occulogram (EOG) and electro-cardiogram. Data were averaged on the onset of each blinks and heart beats separately and PCA was performed separately for each type of sensor. Then, one to three of the first components characterizing the artifact were selected by mean of visual inspection to be further removed.

Data were then entered into Matlab software and processed with Fieldtrip software (http://fieldtrip.fcdonders.nl/). For the first experiment, an automatic rejection of trials based on signal discontinuities (all signal above 30 and 25 standard deviations in 110–140 Hz frequency range) was performed. However, less than 1% of the trials removed, and therefore this step was omitted in experiment 2, where the number of error trials was smaller. A low-pass filter at 30 Hz was then applied as well as a baseline correction from 300 ms to 200 ms before target onset.

Data were then realigned on response onset to be further averaged by subject and conditions. To obtain grand-average evoked response data, we first averaged individual data for each SOA separately, then averaged across SOAs and then across participants. For the first experiment only, response times were equalized across error and correct trials (see Supplementary Methods). Without such a correction, the slower RTs on *seen* correct trials caused artifactual differences due to non-aligned sensory-evoked components on response-locked averages (Fig. S4). This RT correction was not needed in experiment 2 where RTs were longer and response-locked ERPs were therefore uncontaminated by sensory-evoked components. An additional baseline correction was simply performed from 200 to 50 ms before motor response. We verified that these small differences in procedure did not affect the main results, and in particular the same dependency of ERN on visibility was observed when no RT correction was applied to experiment 1 (See Supplementary Results).

Combined EEG/MEG source reconstruction

Brainstorm software was used to derive current estimate from correct and error MEEG waveforms, for each condition of visibility and each subject separately. Cortical surfaces of 22 participants (2 participants were discarded in each experiment as no MRI data could be obtained) were reconstructed from individual MRI with FreeSurfer (http://surfer.nmr.mgh.harvard.edu/) for cortex surface (gray-white matter boundary) and Brainvisa (http://brainvisa.info/) for scalp surface. Inner skull and outer-skull surfaces were estimated by Brainstorm, in order to compute accurate forward model using a three-compartment boundary-element method (OpenMeeg toolbox; http://www-sop.inria.fr/athena/software/OpenMEEG/). Sources were computed with weighted minimum-norm method and dSPM (depth-weighting factor of 0.8, loosing factor of 0.2 for dipole orientation). Individual source estimate data were then projected on a template cortical surface, in order to be averaged across participants, separately for each experiment. Mean power (i.e. square of the t-values) of regions of interest was computed to present time-courses of brain activity.

Statistical analysis

Behavioral data analysis

All behavioral data analyses were performed with Matlab software with the help of the Statistics toolbox using repeated-measures analysis. Reaction-time analysis was performed on the median RT of each condition.

MEG data analysis

To detect significance differences between error and correct conditions for each type of sensor, we used a cluster-based non-parametric t-test with Monte Carlo randomization provided in the Fieldtrip software (Maris and Oostenveld, 2007). This method identifies clusters of nearby sensors presenting a significant difference between two conditions for a sufficient duration while correcting for multiple comparisons. For each sample, t-values and associated p-value were first computed by means of a non-parametric Monte-Carlo randomization test. Clusters were then identified by taking all samples adjacent in space or in time (minimum of 2 sensors per cluster, 4.3 average spatial neighbors per EEG electrode and 8.2 per MEG channel) with $p<0.05$. The final significance of the cluster was found by computing the sum of t-values of the entire cluster, and comparing with the results of Monte-Carlo permutations (1500 permutation). Clusters were considered significant at corrected $p<0.05$ if the probability computed with the Monte-Carlo method was inferior to 2.5% (two-tailed test). Time-windows of interest were chosen for each experiment on the basis of the EEG results for *seen* trials to optimize cluster detectability. The ERN is usually observed in a 100 ms time-window after button press (Dehaene et al., 1994). As the onset of the difference was observed slightly later in experiment 1 than experiment 2, search for clusters was performed respectively on a 30–100 ms time-window after motor response for experiment 1 and 0–100 ms in experiment 2.

For statistical analysis on a-priori clusters, average voltage over central electrodes (FC1, FC2, C1, Cz, C2) were computed over the same time-window as for the cluster analysis (30–100 ms and 0–100 ms after motor response respectively for experiment 1 and 2, analysis of later time windows is reported in Supplementary Results). Analysis was performed in Matlab using repeated-measures t-tests (two-tailed) and ANOVA with visibility and performance as within-subjects factors. Analysis by SOA required more sophisticated statistical analysis as trial rejection and factorial analysis (SOA*Visibility*Performance) led to unequal number of participants in each combination of condition. Therefore, analysis of variance was performed in R software using a

linear mixed-effects model ((Baayen et al., 2008) R package lme4) which allowed us to include all data available (unbalanced design) and still encompass repeated-measures. The functions used yield t statistic and, as degrees of freedom cannot be computed for this kind of analysis, p-values were derived from a Markov Chain Monte Carlo (MCMC) method.

Results

Subjectivity visibility is reliably affected by masking

Subjective visibility, as measured by the percentage of *seen* responses, increased in a non-linear sigmoid manner with SOA ($F_{5,55} = 316.7$, $p < 10^{-4}$, see Supplementary result), replicating earlier results (Del Cul et al., 2007). Stimuli that were masked after a short latency (SOA < ~50 ms) were almost always judged as invisible, while visibility rose very rapidly after this point (Fig. 2). Visibility was slightly higher in experiment 1 compared to experiment 2 (two way ANOVA with factor experiment and SOA, $F_{1,55} = 3.371$, $p = 0.094$), probably because participants underwent more training in experiment 1 than in experiment 2. However, the main effect of SOA was highly significant in both cases, and no interaction was found between SOA and experiment ($F_{5,55} = 1.77$, $p = 0.135$).

Raw visibility reports (*Seen*, *Unseen*) can be criticized as subjective and potentially biased measures. We therefore transformed them into an objective index of target detection sensitivity and bias, using classical signal detection theory. To this end, at each SOA level, visibility ratings (percent *Seen* responses) were compared against those in the mask-only condition, and converted to *detection-d′* and bias values (see Materials & methods). For the shortest SOA condition (SOA = 16 ms), participants were at chance to detect the presence of the target, as the *detection-d′* did not differ significantly from 0 (Exp1: average $d' = 0.15$, $t_{12} = 0.98$, $p = 0.34$, Exp2: average $d' = 0.01$, $t_{12} = 0.07$, $p = 0.94$). Furthermore, participants adopted a conservative criterion (bias > 0, $t_{12} = 14.6$, $p < 10^{-4}$, $t_{12} = 17$, $p < 10^{-4}$), reflecting the frequent use of the *unseen* response on both target-present and mask-only trials, and therefore confirming the invisibility of the targets at this SOA. As SOA increased, *detection-d′* increased ($F_{4,44} = 220.7$, $p < 10^{-4}$) while response-bias toward the *unseen* response decreased ($F_{4,44} = 221$, $p < 10^{-4}$), confirming that visibility improved with SOA. Finally, on mask-only trials, false-positives were very rare (exp 1: 3% erroneous *seen* responses; exp 2: 4%). Overall, these observations confirm that subjective visibility reports were reliable and that masking at short SOA induced a subjective state of invisibility on a large proportion of trials.

Cognitive and metacognitive performance are affected by masking

We then looked at the variations in performance and meta-performance as a function of SOA (see Fig. 2; Response times are reported in Supplementary material).

Objective performance in the number comparison task increased with SOA ($F_{4,44} = 318.89$, $p < 10^{-4}$), with a non-linear profile virtually parallel to subjective visibility (Figs. 2C-D). As intended, in the first experiment where strong time pressure was imposed, participant's performance did not reach ceiling even for the largest SOA (SOA

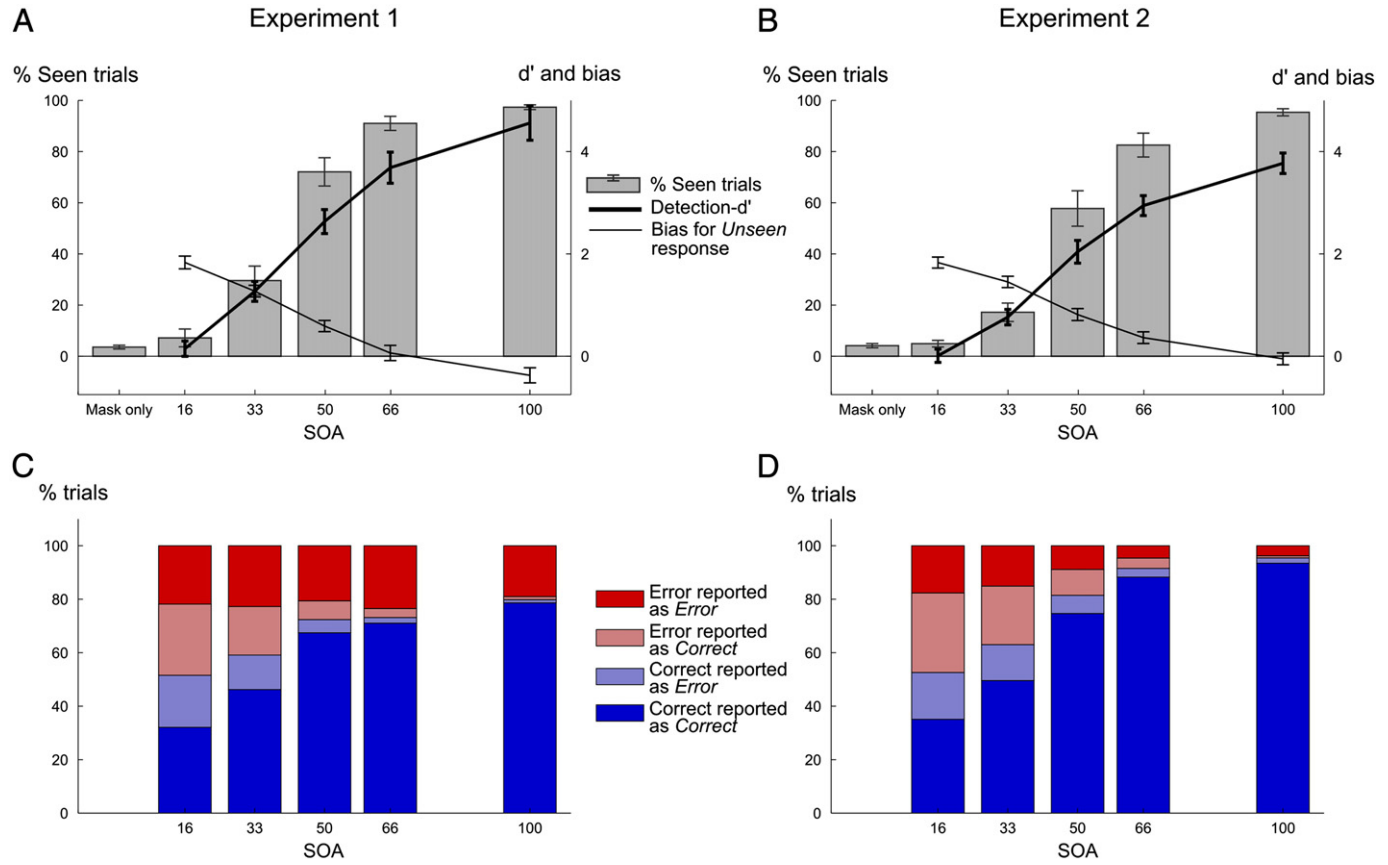


Fig. 2. Visibility and performance results according to SOA for experiment 1 (left column) and 2 (right column). (A–B) Visibility ratings, expressed as the proportion of *seen* responses (left axis ranging from 0 to 100%) as a function of SOA. The thick line represents *detection-d′* values (right axis, ranging from 0 to 4) while the thin line represents response bias towards *unseen* response (same scale as *detection-d′*), for each SOA. (C-D) Percentage of each category of trials according to actual objective performance and subjective report of performance (Error trials correctly classified as Error in dark red, Correct trials correctly classified as Correct in dark blue, Error trials incorrectly classified as Correct in light red and Correct trials incorrectly classified as Error in light blue), for each SOA.

100 ms, Fig. 2C). Thus, experiment 1 achieved its goal of generating a minimum of ~20% errors at each SOA, allowing us to explore the mechanisms of error detection. In the second experiment, where time pressure was relaxed, performance at the longest SOA reached 95% correct (Fig. 2D), thus resulting in a much smaller number of analyzable errors. This pattern resulted in a significant SOA by experiment interaction ($F_{4,44} = 19.49$, $p < 10^{-4}$).

Next, we investigated meta-cognitive performance as a function of SOA. Our procedure allowed us to compare, on each trial, the subject's objective accuracy with his evaluation of his performance. Trials were classified as "*meta-correct*" if they were error trials perceived as errors, or correct trials perceived as correct. Otherwise they were labelled as "*meta-incorrect*". Meta-cognitive performance (i.e. percentage of meta-correct trials) increased with SOA ($F_{4,44} = 165.83$, $p < 10^{-4}$), reaching 97% meta-correct trials in both experiments. As seen on Figs. 2C–D, both types of meta-incorrect responses (undetected errors as well as correct trials misperceived as errors) progressively vanished with increasing SOA, in tight parallel with increasing target visibility.

Overall, these results indicate that the SOA manipulation successfully modulated, in tight parallel, the performance of our three tasks: objective number comparison, metacognitive evaluation, and visibility judgment. In the next section, we show how visibility, independently of SOA, indexes a major switch in the performance of the other two tasks.

Cognitive and metacognitive performance are affected by visibility

To better characterize how behavior changed on conscious and non-conscious trials, the data were then split by visibility (*Seen* vs *Unseen*). As visibility increased in a non-linear way with SOA, many participants had fewer than 5 trials in one of the visibility condition for extreme SOA values. Therefore, we removed these trials from the analysis and from the figures, keeping for *seen* trials only trials corresponding to SOA larger than 33 ms and for *unseen* trials those corresponding to SOA smaller than 50 ms.

As can be seen in Figs. 3A–B, participants performed way above chance both in the number comparison task and in the performance evaluation task when they could see the target number, independently of the SOA condition (for experiments and all SOA, performance and meta-performance > 50%, $p < 0.005$). When averaging together all SOAs or when considering only intermediate SOAs (33 and 50 ms) for which we had approximately as many *seen* and *unseen* trials, both performance and meta-performance were significantly superior on *seen* compared to *unseen* trials (for both experiments, all $p < 0.01$). This finding was similar in both experiments, with a small difference: for the *seen* trials, at the longest SOA (100 ms), performance was lower in experiment 1 compared to experiment 2 (80% versus 96%), again because of the strong time pressure imposed in experiment 1.

To obtain a clearer view of the relative sensitivity of the subject in the second-order performance evaluation task compared to the primary task, performance was converted to *d′* and *meta-d′* values (Figs. 3C–D). As described by second-order Signal Detection Theory (Galvin et al., 2003; Maniscalco and Lau, 2012; Rounis et al., 2010) (SDT), *d′* and *meta-d′* give an unbiased estimate of performance, respectively for first-order task (here, number comparison) and second-order task (error detection). Since these two measures are on the same scale, they allow us to compare what the first-order performance actually was to what it should have been, given second-order error detection accuracy (Galvin et al., 2003; Maniscalco and Lau, 2012; Rounis et al., 2010).

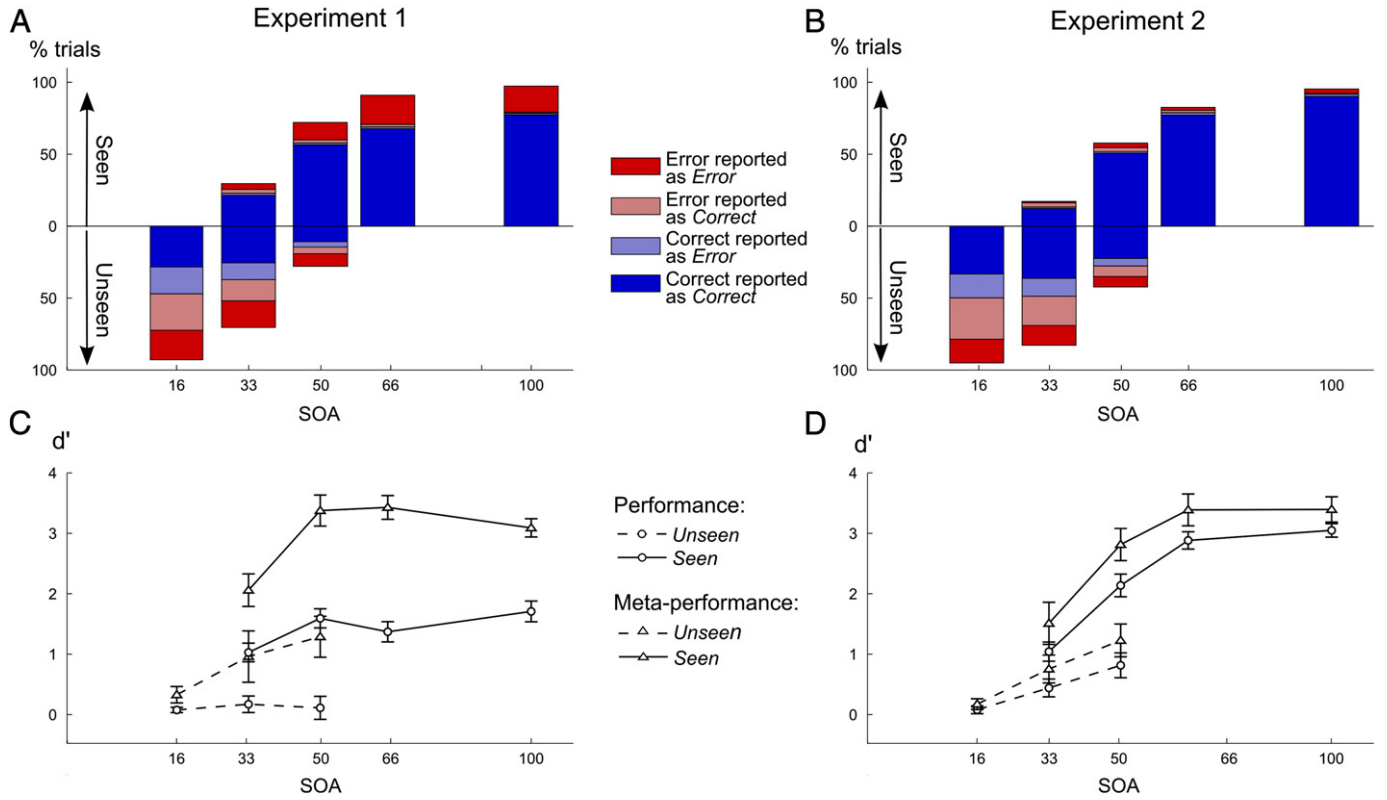


Fig. 3. Performance and meta-performance according to visibility and SOA in both experiments (left column, experiment 1; right column, experiment 2). (A–B) Proportions of *unseen* (below midline) and *seen* trials (above midline) were computed for each SOA. For each type of trials and each SOA, the relative percentage of each category of trials was derived according to objective performance and subjective report of performance (same color code as in Fig. 2). (C-D) Unbiased measures of performance (*d′*, circles) and meta-performance (*meta-d′*, triangles) were computed separately for *seen* (solid line) and *unseen* (dashed-line) trials and each SOA value. All error-bars represent standard error.

This analysis confirmed that even for equal SOA, both performance and meta-performance showed a sudden jump with visibility (see Figs. 3C–D; statistics in Table 1). Thus, visibility judgment, although a subjective task, also indexes a large change in objective performance: *seen* and *unseen* trials differ massively in the quantity of usable information for both primary and secondary judgments (Del Cul et al., 2007, 2009).

For *seen* trials (Figs. 3C-D, solid lines), performance and meta-performance (d' and *meta-d′*) increased significantly with SOA in both experiments (see Table 2). *Meta-d′* always significantly exceeded d', in particular in Experiment 1 with time pressure ($F_{1,12}=167.3$, $p<10^{-4}$), but also in Experiment 2 ($F_{1,12}=9.93$, $p=0.008$). This finding indicates that some of the primary responses were errors that could be detected prior to second-order judgment, resulting in "change-of-mind" (Resulaj et al., 2009). In sum, on *seen* trials, participants managed to perform the metacognitive task with very high accuracy.

Cognitive and metacognitive performance are above chance on unseen trials

We next performed similar analyses of cognitive and metacognitive performance restricted to the *unseen* trials.

For first-order performance, performance remained at chance level on *unseen* trials in experiment 1 (%correct = 50%, for all SOA, $p>0.30$, Fig. 3A), presumably due to the pressure on speed. In experiment 2, when time pressure was relaxed, performance slightly surpassed 50% (%correct > 50%, for all SOA, $p<0.05$, Fig. 3B).

These results were confirmed by an analysis of first-order d' values. In experiment 1, performance was at chance for all SOAs ($d'=0$, all $p>0.10$, Fig. 3C), but once speed pressure was relaxed in experiment 2 (Fig. 3D), objective performance increased with SOA ($F_{2,24}=10.589$, $p=0.0005$) and differed from chance for SOA 33 ms ($t_{12}=2.99$, $p=0.011$) and 50 ms ($t_{12}=3.97$, $p=0.002$). Experiment 2 thus demonstrates a classical subliminal effect (Persaud et al., 2007; Pessiglione et al., 2007), i.e. a partial accumulation of evidence about the *unseen* targets.

Most importantly, second-order performance in the error detection task (i.e. meta-performance) was significantly above chance in both experiments for intermediate SOAs (SOA 33 and 50 ms, meta-performance > 50%, all $p<0.005$). Indeed, as shown in Figs. 3A–B, when pooling these two intermediate SOAs, a large number of correct trials were correctly classified as such (exp 1: 65.8%; exp 2: 72.9%). Again, SDT analysis confirmed this result, as *meta-d′* was significantly superior to 0 (chance level) on *unseen* trials, both in experiment 1 (SOA 16 ms: $t_{12}=2.42$, $p=0.032$, SOA 33 ms: $t_{12}=2.26$, $p=0.043$ and SOA 50 ms: $t_{12}=3.79$, $p=0.003$) and in experiment 2 (SOA 33 ms: $t_{12}=3.27$, $p=0.007$ and SOA 50 ms: $t_{12}=4.52$, $p=0.0007$) and seem to increase with SOA (Exp1: $F_{2,24}=2.65$, $p=0.091$; $F_{2,24}=8.50$, $p=0.002$).

Direct comparison of d' and *meta-d′* showed that, for both experiments, meta-cognitive performance exceeded primary task performance on *unseen* trials. This was true over all *unseen* trials (SOA 16–50 ms, Exp1: $F_{1,60}=11.48$, $p=0.005$; Exp 2: $F_{1,60}=13.2$, $p=0.003$), at intermediate SOAs 33 ms (Exp1: $t_{12}=-1.89$, $p=0.041$; Exp2: $t_{12}=-1.97$, $p=0.036$) and at SOA 50 ms (Exp1: $t_{12}=-3.28$, $p=0.003$; Exp2: $t_{12}=-2.09$, $p=0.023$). Even in subliminal conditions, once a primary response is emitted, participants can categorize it as correct or incorrect with better-than-chance performance.

Table 2
Statistical increase in performance and meta-performance with SOA for experiment 1 and 2.

	Experiment 1	Experiment 2
d'	$F_{3,36}=8.776$, $p=0.0002$	$F_{3,36}=49.677$, $p<10^{-4}$
meta-d′	$F_{3,36}=8.12$, $p=0.0003$	$F_{3,36}=10.3$, $p<10^{-4}$

To summarize, we found that in both experiments, participants were above chance in judging their own errors, even on trials classified as *unseen*. Most remarkably, for subliminal stimuli in experiment 1, participants were at chance for the objective task, presumably due to time pressure, and yet they were still able to evaluate their accuracy better than chance. In experiment 2, they were above chance for both cognitive and metacognitive tasks, a result that may relate to the reduced time pressure compared to experiment 1.

The error-related negativity is present only on seen trials

We then turned to EEG recordings, in order to probe whether metacognitive performance was accompanied by an ERN, even under subliminal conditions (Fig. 4).

Starting with the *seen* trials, a significant ERN, manifested by more negative central voltages on error than on correct trials, was found in both experiments (Figs. 4A-B, Exp. 1: $t_{12}=-3.39$, $p=0.0053$; Experiment 2: $t_{12}=-3.42$, $p=0.0051$). Importantly, no significant difference was detectable on *unseen* trials in experiment 1 ($t_{12}=-0.55$ $p=0.59$), suggesting that the ERN was absent under subliminal conditions. In this experiment, the number-comparison task was strongly speeded, leaving open the possibility that the results might be an artefact of time–pressure, with the response being emitted too fast to observe an ERN. However, this interpretation was rejected by experiment 2, where a similar result was observed ($t_{12}=0.02$, $p=0.98$) although time–pressure was relaxed and response-time was longer (see Supplementary material).

The variation of the ERN with subjective report was confirmed by a significant interaction between visibility (*seen* or *unseen*) and performance (*error* or *correct*) on central voltages in the time window of the ERN (Exp 1 $F_{1,36}=8.62$, $p=0.012$; Exp 2 $F_{1,36}=10.46$, $p=0.0072$, see Materials & methods). The ERN remained undetectable on *unseen* trials, even when we restricted the analysis to trials in which metacognitive performance was correct (see Supplementary Results) and therefore a maximal amount of stimulus information was accumulated. The absence of the ERN on these trials suggests that above-chance metacognitive performance on subliminal trials was not mediated by the ERN, which was simply absent or drastically reduced under subliminal conditions.

The ERN depends on visibility, not SOA

The above *seen/unseen* comparison is partially confounded with differences in SOA, as the majority of *seen* trials comes from trials with long SOAs. It could therefore be argued that the presence of the ERN on *seen* trials has nothing to do with subjective visibility, but is simply due to the additional information made available by

Table 1
Statistical analyses of performance and meta-performance scores, relative to chance level, as a function of visibility, for experiment 1 and 2.

		Pooling all SOAs		SOA 33 ms		SOA 50 ms	
Performance	exp 1	$t_{12}=10.5$	$p<10^{-4}$	$t_{12}=5.20$	$p<10^{-4}$	$t_{12}=6.9921$	$p<10^{-4}$
	exp 2	$t_{12}=12.5$	$p<10^{-4}$	$t_{12}=3.70$	$p=0.0015$	$t_{12}=5.08$	$p=0.0001$
Meta-performance	exp 1	$t_{12}=9.42$	$p<10^{-4}$	$t_{12}=2.719$	$p=0.0093$	$t_{12}=4.507$	$p=0.0003$
	exp 2	$t_{12}=8.73$	$p<10^{-4}$	$t_{12}=1.677$	$p=0.0597$	$t_{12}=5.15$	$p=0.0001$

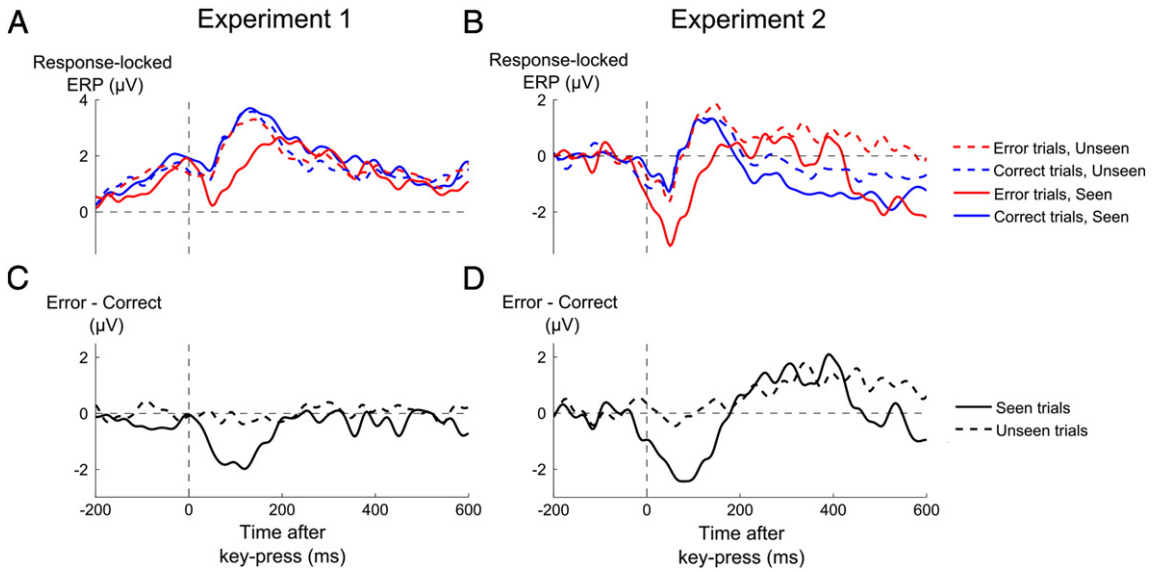


Fig. 4. Time courses of event-related potentials as a function of objective performance and visibility. (A,B) Grand-average event-related potentials (ERPs) recorded from a cluster of central electrodes (FC1, FC2, C1, Cz, C2), sorted as a function of whether performance was erroneous (red lines) or correct (blue lines), and whether the target was *seen* (solid lines) or *unseen* trials (dashed lines), for experiment 1 (A) and experiment 2 (B). (C,D) Difference waveforms of error minus correct trials, separately for *seen* (solid line) and *unseen* (dashed line) trials.

the longer SOA (indeed, a similar confound applies to previous research by Pavone et al. (2009) and Woodman (2010). However, because we collected visibility information on every trial, our design allowed bypassing this limitation. We sorted the trials as a function of both SOA and trial-by-trial judgement of visibility, taking advantage of spontaneous fluctuations in visibility for a fixed SOA. This analysis could only be performed in experiment 1 as too few error trials occurred in experiment 2.

On *unseen* trials, a general linear model (see Materials & methods) with SOA (16, 33 or 50 ms) and performance (*correct* or *error*) as within-subject factors confirmed the absence of a difference between error and correct trials (no ERN, $p = 0.91$, Fig. 5F) and no interaction with SOA ($p = 0.76$). Indeed, none of the SOAs showed a significant ERN (all $p > 0.25$). For *seen* trials, conversely, a similar ANOVA over SOAs 33, 50, 66 and 100 ms revealed a main difference between error and correct trials ($p < 10^{-4}$, Fig. 5E). Furthermore, an interaction with SOA ($p = 0.04$) indicated that the ERN increased with SOA.

Most crucially, for SOA 50 ms, the voltage difference between correct and error trials varied drastically with visibility. No ERN was observed for *unseen* trials ($t_{10} = 0.58$, $p = 0.29$, Fig. 5F) while a clear ERN

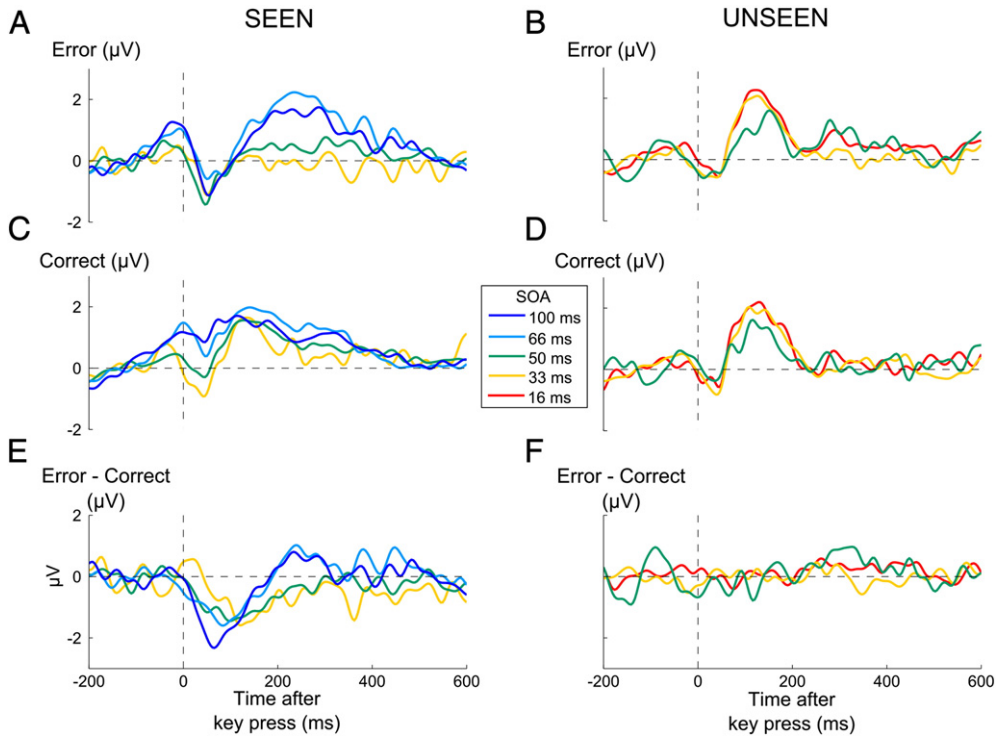


Fig. 5. Time courses of event-related potentials as a function of SOA and objective performance for *seen* and *unseen* trials. (A–D) Grand-average event-related potentials (ERPs) by SOA condition for error (top raw, A and B) and correct (middle raw, C and D) trials in *seen* (left column, A and C) and *unseen* (right column, B and D) conditions for experiment 1 on a cluster of central electrodes (FC1, FC2, C1, Cz, C2). (E,F) Difference waveforms of error minus correct for *seen* (solid line) and *unseen* (dashed line) trials, by SOA. Due to reduced trial numbers, only the shortest SOA (16, 33 and 50) ms are presented for *unseen* trials while only longer SOAs (33 ms, 50 ms, 66 ms and 100 ms) are included for *seen* trials.

was present for *seen* trials ($t_{11} = 2.48$ $p = 0.015$, Fig. 5E). Thus, subjective visibility, over and above objective variations in SOA, determined the presence or absence of an ERN. For SOA 33 ms, the difference between error and correct trials did not reach significance neither for the *unseen* ($t_{12} = -0.23$, $p = 0.59$), nor for the *seen* trials ($t_8 = 1.16$, $p = 0.14$) probably due to the small number of participants having enough data points in this condition. Fig. 5E suggests that at this SOA, the ERN was present but temporally spread out, which we verified by observing significantly more negative voltages for errors than for correct trials once averaging over the interval 50-200 ms ($t_8 = 2.53$, $p = 0.018$). Within the *seen* trials, the error-correct difference reached significance for all other SOAs (SOA 66 ms: $t_{11} = 3.02$, $p = 0.006$; SOA 100 ms: $t_{11} = 3.37$, $p = 0.003$).

In summary, at any SOA, the ERN was present if and only if participants reported seeing the target.

MEG detects signatures of conscious and non-conscious errors

To identify the cerebral signatures of error processing, cluster analysis was applied to MEG and EEG data in order to identify any cluster of sensors showing a difference between error and correct trials. To take advantage of the possible differences in sensitivity between sensors, we analyzed separately each type of sensor (electrodes, magnetometers, longitudinal and latitudinal gradiometers) for *seen* and *unseen* trials. For EEG, cluster analysis essentially replicated the above ERN analysis. On *seen* trials, a significant cluster, with more negative voltages on error trials, was found on fronto-central electrodes in EEG, for both experiment 1 ($p = 0.0067$, Fig. 6A) and 2 ($p = 0.0013$, Fig. 6C). The cluster began at motor onset in experiment 2, and continued for 100 ms, while it started at 50 ms after the response in experiment 1. In *unseen* trials, no significant EEG cluster was detected.

For MEG, in experiment 1, significant clusters were found for two of the three types of channels in the *seen* trials (Fig. 6A, latitudinal gradiometers cluster: left fronto-lateral region, 25–70 ms after response, $p = 0.015$; magnetometers cluster: right parieto-central region, 65–90 ms, $p = 0.023$), suggesting different sensitivity to error-related signals across sensor types. Again however, no significant cluster was found for the *unseen* trials (Fig. 6B).

As time–pressure induced speeded responses in experiment 1, we then turned to experiment 2, in which more evidence should be available at response onset and error-related processes should have full ability to develop. Indeed, MEG sensors revealed a different pattern of activity for this experiment. For *seen* trials, only magnetometers (Fig. 6C) showed error-related activity (orbito to dorso-frontal regions, 5–55 ms). More surprisingly, even for *unseen* trials, significant differences were observed in two clusters of sensors (Fig. 6D; longitudinal gradiometers, 0–65 ms, $p = 0.002$; magnetometers, 0–45 ms, $p = 0.007$), none of them resembling however with those found for the *seen* trials. These results suggest that MEG sensors may provide a more sensitive and comprehensive view of error-processes than EEG, a result that is coherent with recent studies showing accrued sensitivity of MEG sensors to sources located in the cingulate gyrus, where the generators of the ERN are thought to be located

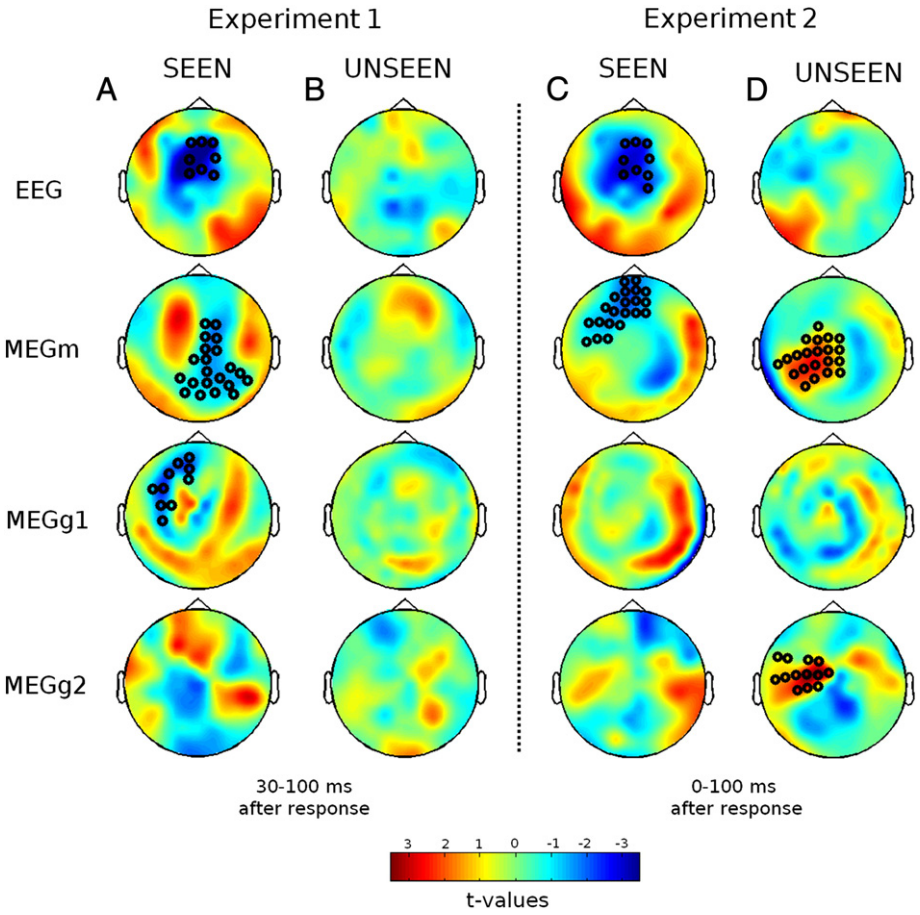


Fig. 6. Error-related MEEG topographies as a function of target visibility. Each plot depicts the scalp topography of the t-value for a difference between correct and error trials, averaged across a 30–100 ms time window for experiment 1 and 0–100 ms for experiment 2 following the motor response, separately for each type of sensors (EEG, magnetometers [MEGm], longitudinal gradiometers [MEGg1], latitudinal gradiometers [MEGg2]) and for the *seen* and *unseen* trials, in experiments 1 (A) and 2 (B). Black circles indicate sensors belonging to a spatiotemporal cluster showing a significant difference ($p < 0.025$) between error and correct conditions using a Monte-Carlo permutation test.

(Irimia et al., 2011). Furthermore, this analysis confirms that these error-processes are modulated by consciousness but also by time–pressure as different results were obtained in the two experiments.

Conscious error detection originates from posterior cingulate cortex

To shed more light on the cerebral generators of these error responses observed at the sensor level, we applied distributed source estimation on error and correct MEEG signals. For *seen* trials in experiment 1, the main source of the difference between error and correct trials was found bilaterally in the anterior part of the Posterior Cingulate Cortex (PCC, Fig. 7A). Its time course matched the dynamics of the ERN (Fig. 7E), and its peak coordinates (Talairach coordinates $x=-6$ $y=-22$ $z=33$) felt close to a recently published MEEG and fMRI study (Agam et al., 2011). In the *unseen* condition, this activity was drastically reduced, in accordance with the absence of a significant effect at the sensor level. Lowering the threshold only revealed weak and inconsistent differences in the most posterior part of the cingulate cortex (Fig. 7C).

In experiment 2, the involvement of PCC on conscious errors was replicated (Talairach coordinates $x=-9$ $y=-23$ $z=31$), but additional error-related activity was also observed in dorsal anterior cingulate (dACC, Talairach peak at coordinates $x=7$ $y=2$ $z=27$, Figs. 7B and F), explaining the observed differences in MEG sensor-level topographies in experiments 1 versus 2. Again, activation in these regions was drastically reduced for *unseen* trials. Nevertheless, small patches in dACC (Fig. 7D) remained active in the *unseen* condition, compatible with the small but significant effect detected at the sensor level in MEG data.

When further restricting the analysis to *unseen* meta-correct trials, in which performance was correctly evaluated (see Supplementary Results), time-courses indeed revealed a short-lived response (Fig. S5) in dACC coinciding with the early part of the error-related activation observed on *seen* trials. Thus, this transient dACC activation might be one of the substrates for above-chance metacognitive performance.

Discussion

In this study we explored whether the meta-cognitive process of error detection in a simple response-time decision task requires conscious perception of the stimulus in order to be deployed. We recorded brain responses in a masking paradigm with variable time–pressure and masking strength, and evaluated the relation between first-order performance, meta-cognition, and subjective visibility. Our findings indicate that two types of metacognitive processes have to be distinguished: (1) The likelihood of having made an error can be estimated above chance level, in a statistical manner, even when making a forced-choice response to a subliminal stimulus; (2) the ERN, which reflects the detection of whether an error was made on a given trial, indexes another process that is only deployed on trials where the stimulus is consciously perceived.

Metacognition without consciousness

Behaviorally, we compared performance in the number comparison task and in the meta-performance task of detecting one's own errors. For the latter, following Maniscalco and Lau (2012), we used a *meta-d′* measure that evaluates what should have been the performance in the first-order task given the performance observed in the second order task. This method allowed us to compare, on the same scale, performance in the number comparison task (*d′*) and performance in error detection (*meta-d′*).

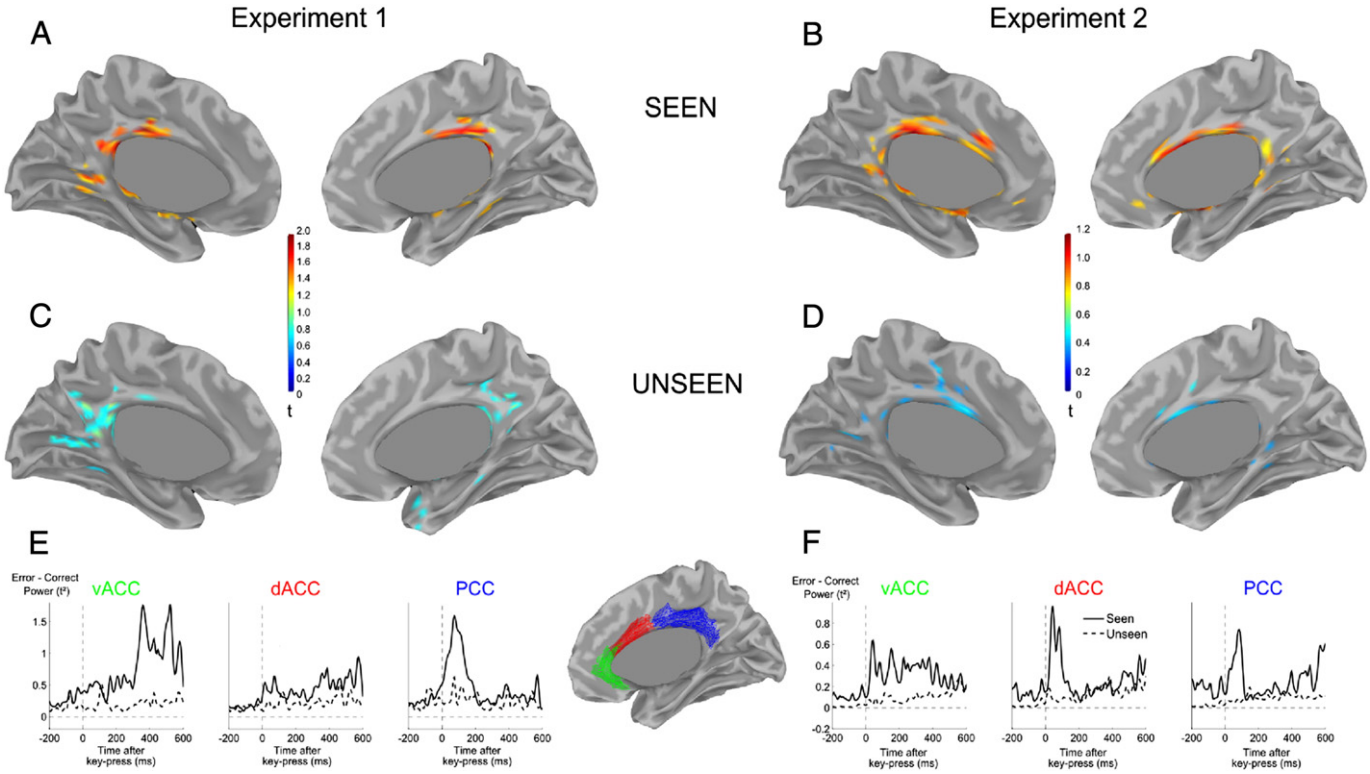


Fig. 7. Difference of source estimates between error and correct MEEG signals. (A–D) View of the medial surface of the left and right hemispheres, for experiment 1 (A,C) and experiment 2 (B,D), for *seen* (A–B) and *unseen* (C–D) trials. Data are thresholded at 66% of maximum activity within each condition. Brain activity was averaged in a 30–100 ms time-window for experiment 1 (A,C) and 0–100 ms for experiment 2 (B,D). (E–F) Time-courses of brain activity in three bilateral regions of interest located in ventral Anterior Cingulate Cortex (vACC), dorsal Anterior Cingulate Cortex (dACC) and Posterior Cingulate Cortex (PCC), for experiment 1 (E) and experiment (2), for *seen* (solid-line) and *unseen* (dashed-line) trials. Values correspond to instantaneous power in the region of interest (average, across vertices, of the square current density t-maps).

In two distinct experiments, we found that participants were able to do better than chance in detecting their own performance under conscious, but also under non-conscious conditions. In Experiment 1, meta-performance in error detection exceeded performance in the first-order task, presumably because, under time–pressure, the primary response was emitted too early, and participants later revised their judgments using a more complete accumulation of evidence on the stimulus (Resulaj et al., 2009). This interpretation was supported by Experiment 2: when time–pressure was weakened, both performance and meta-performance reached above-chance levels and evolved in close parallel as a function of SOA (Fig. 3).

Crucially, participants performed above chance in detecting their own errors even on *unseen* trials. In both experiments, meta-cognitive performance on *unseen* trials increased with SOA, suggesting that longer SOAs allowed increasing amounts of evidence to be accumulated, as previously demonstrated for subliminal visual and motor processing (Del Cul et al., 2007; Vorberg and Mattler, 2003).

Our findings therefore suggest that meta-cognition should be added to the list of processes that can be partially deployed non-consciously. Such a result is in line with a previous report showing a higher-than-chance performance in metacognitive judgments of confidence under conditions of invisibility due to inattention (Kanai et al., 2010). Similarly, another study showed that a blindsight patient was able to perform above chance-level in his second-order confidence judgments, even when the stimulus was presented in his blind hemi-field (Evans and Azzopardi, 2007). Such findings contradict the view that under conditions of subjective invisibility, participants are not able to predict their accuracy in detecting a masked target. Indeed, measurement of post-error slowing suggests that participants are able to monitor their performance non-consciously, and are sensitive to their objective errors even when the experimental paradigm misleads them into thinking that their performance was correct (Logan and Crump, 2010).

These findings conflict with the common intuition according to which self-oriented monitoring processes are tightly linked to consciousness (Kunimoto et al., 2001; Lau and Passingham, 2006; Persaud et al., 2007). In particular, our finding that above-chance metacognitive judgments do not necessarily indicate conscious perception of the stimulus seems incompatible with the use of wagering or confidence as an index of consciousness (Kunimoto et al., 2001; Persaud et al., 2007). Nonetheless, such a critique must be qualified, as above-chance subliminal metacognition is probably limited to experimental circumstances where a forced-choice judgment is imposed. Furthermore, in the present study, participants had to be explicitly informed that even when responding randomly they still had a 50% chance of being correct. Therefore they should venture "error" and "correct" responses on approximately half of trials. Prior to this instruction, a pilot study showed that most of them spontaneously responded with the "error" key on all unseen trials, suggesting a total lack of confidence in their capacity to make both first- and second-error judgments. In the same manner, blindsight patients may first have to gain an explicit awareness that their performance largely exceeds chance level before performing a second-order metacognitive task (Evans and Azzopardi, 2007). It remains unclear whether above-chance subliminal metacognitive abilities would be observed without this prior knowledge of first-order accuracy. In that sense, wagering and confidence judgments may vary more tightly with subjective reports of visibility in some contexts than others. Altogether however, these findings confirm that, as any other decision processes, second-order judgments are subject to response biases (Evans and Azzopardi, 2007; Fleming and Dolan, 2010) and should therefore be analyzed carefully to disentangle the effect of criterion setting from the true level of "meta-evidence" available about a given cognitive process.

Second-order signal detection theory (SDT) offers a theoretical framework within which to analyze such measures, and is capable of explaining both first- and second-order non-conscious performance. According to classical SDT, an observer receives a sensory sample on a continuum, and the first-order response is selected by deciding on which side of a decision boundary it falls. Second-order SDT points out that information on the distance of the sensory evidence from the decision boundary can be used to partially predict response accuracy, thus supporting a second-order judgement (Galvin et al., 2003). Intuitively, sensory evidence that falls very close to the decision boundary is highly ambiguous and will therefore likely lead to an error. In contrast, sensory evidence that falls far from the boundary is (statistically) more indicative of a correct response. According to this model, decision and confidence are therefore computed simultaneously from the same data. Previous behavioral and neural evidence (Kepecs et al., 2008; Kiani and Shadlen, 2009; Resulaj et al., 2009) supports this view. Furthermore, the theory can explain the gist of our present results: since first-order evidence towards a decision can be accumulated from *unseen* stimuli, resulting in above-chance first-order performance (Vorberg and Mattler, 2003), it follows from the theory that it should also be possible for the same system to compute second-order confidence information non-consciously — as demonstrated here.

However, the data of Experiment 1 impose a small revision on the second-order SDT mechanism proposed by Galvin et al. (2003). This theory supposes that a single sample of sensory evidence is used for both first-order and second-order tasks, predicting that meta-performance cannot exceed performance (Galvin et al., 2003). However, in Experiment 1, under strong time pressure, primary judgment was at chance while second-order performance was above chance. In that respect, our findings are reminiscent of the observation of "changes-of-mind" in a sensori-motor task, i.e. accurate corrective movements performed after the first response was launched even though no additional sensory data was provided (Resulaj et al., 2009). Both findings can be accounted for by supposing that early responses do not fully make use of the available sensory evidence and that, with additional time, participants can accumulate additional evidence in order to ultimately revise their judgments. Indeed, when we removed time pressure in Experiment 2, both performance and meta-performance became aligned with each other (d' and *meta-d′* did not differ).

The SDT framework can be modified to take into account such dynamics of decision making (Resulaj et al., 2009). Indeed, the recently introduced Two-Stage Dynamic Signal Detection Theory (Pleskac and Busemeyer, 2010) integrates these two elements into a framework that accurately predicts both the dynamics of decision-making and subsequent confidence judgments. This model allows additional processing of the stimulus to take place even after an initial decision has been made. Such feature results in confidence judgments that can potentially rely on more information than primary choices, especially when speed is emphasized over accuracy, exactly as observed in our study.

All-or-none error detection and conscious perception

The SDT framework for metacognition is, however, inherently limited. It is continuous and statistical in nature, and cannot label, with near-certainty, whether a given trial was correct or erroneous. Rather, it merely achieves above-chance meta-performance on average. While such a statistical mechanism adequately accounts for the observed metacognitive performance on subliminal trials, it seems insufficient to explain error detection on conscious trials. When participants reported seeing the stimuli, they were often highly confident in the detection of their errors, and accurately categorized their performance on each trial in the absence of any feedback (Fig. 3). A distinct mechanism therefore seems needed to account for the capacity to label specific trials as erroneous, which only occurred on conscious trials. Indeed, EEG and MEG recordings gave evidence that a distinct performance monitoring mechanism, indexed by the ERN, was deployed exclusively on conscious trials.

In Experiment 1, the ERN was detectable on conscious trials but was drastically reduced to undetectable levels when participants reported

not seeing the target. This result was confirmed by an analysis of the neural generators of the ERN, whose activation showed a step-like increase with visibility. Even for identical masking strength, the ERN was observed on *seen* trials but not on *unseen* trials. This result was replicated in Experiment 2 where the pressure to respond quickly was removed, showing that the absence of a subliminal ERN was not caused by a lack of processing time.

Our results replicate and extend prior research using a 4-dot masking task (Woodman, 2010). In this task, Woodman observed an ERN when the target was consciously perceived, but not when it was masked and became invisible. In this study, however, visibility was confounded with a physical change in the display (delayed mask offset). Our study goes beyond their finding by taking advantage of the spontaneous fluctuations in visibility that occur for a fixed stimulus. We demonstrate that the ERN is modulated purely as function of subjective reportability without any objective change in the stimulus. Our study also shows that the absence of the ERN needs not be accompanied by a lack of meta-cognitive performance, and provides information as to the generators of these two error monitoring devices.

In contrast to the results of Woodman (2010), Pavone et al. (2009) reported the detection of a significant ERN on both unaware and aware errors, compared to correct trials. A close examination of their graphs, however, suggests that their difference might be related to pre-response baseline shifts, possibly due to the fact that response times were not equalized. Note that in our experiment, we only examined the ERPs to error and correct trials that were carefully equalized to have equal distributions of responses times (see Materials & methods). A failure to do so may result in the emergence of artifactual differences in the time course of the ERPs which are unrelated to errors themselves, but simply reflect variations in response speed between correct and error trials. If a baseline correction was applied to Pavone et al.'s results, their graphs suggest that an identical negativity would be seen on correct and erroneous subliminal trials — i.e. an absence of a subliminal ERN, similar to what we observed.

Some studies aimed at manipulating more directly the awareness of making an error which, as we noted in the Introduction, constitutes a different question. In antisaccade studies (Endrass et al., 2007; Nieuwenhuis et al., 2001; Wessel et al., 2011) an ERN has been observed when participants made eye-movement errors that were not consciously detected. The apparent conflict with our work is only superficial as in these studies the target was always consciously visible and a conscious motor intention could always be prepared. The only aspect of which participants remained unaware was the deviation of their actual movements from the intended trajectory. Their results therefore suggest that the ERN may remain present when the action itself is non-conscious. In contrast, our results suggest that the ERN vanishes when the target, and therefore the correct response, cannot be consciously represented.

Other studies (Dhar et al., 2011; Hughes and Yeung, 2011; O'Connell et al., 2007), focused exclusively on error awareness in experimental paradigms where conflicting stimulus–response rules induced confusions on the nature of the correct response. Again, they found that the ERN was present even for errors that were undetected. However it remains unclear in such paradigms whether participants were unaware of their errors because of an erroneous representation of the correct response, or because of a failure in the error-detection process itself. In either case, such results do not conflict with our finding as these studies did not manipulate awareness of the stimulus itself but rather introduced confusion on the stimulus–response mapping.

A converging finding of these studies, confirmed by others (Hewig et al., 2011; Hughes and Yeung, 2011; Steinhauser and Yeung, 2010), is that the ERN does not necessarily signal a consciously perceived error. Again, this conclusion is not incompatible with our result: while the ERN is evoked only when a conscious target is present, it may not yet reflect the conscious detection of the error. Rather, it may just index an intermediate process on the way to conscious error detection. Indeed, several recent articles suggest that error awareness might be related to the error positivity (Pe) (Dhar et al., 2011; Endrass et al., 2007; Hewig et al., 2011; Hughes and Yeung, 2011; Nieuwenhuis et al., 2001; O'Connell et al., 2007; Steinhauser and Yeung, 2010) which follows the ERN. In that sense, the Pe may be analogous to the sensory P3 potential observed in many experiments where conscious and unconscious sensory trials are contrasted (Dehaene and Changeux, 2011). A detailed analysis of the behavior of the Pe in our two experiments, confirming the dissociation between ERN and Pe and partially supporting the above hypotheses, may be found in Supplementary materials (see also Fig. 4).

The present results further clarify the types of brain events that occur when a sensory stimulus becomes conscious and crosses the threshold for reportability. The Global Neuronal Workspace (GNW) model proposes that conscious access is associated with a sharp non-linear transition in brain activity (Dehaene and Changeux, 2011), leading to an all-or-none change in subjective reports and late brain activity on *seen* compared to *unseen* trials (Del Cul et al., 2007; Quiroga et al., 2008; Sergent and Dehaene, 2004b; Sergent et al., 2005). However, this all-or-none view has been challenged on the grounds that behavioral measures, priming, and brain activation often show a continuous rather than discontinuous reduction on subliminal relative to supraliminal trials (Dehaene et al., 1998; Overgaard et al., 2006; van Gaal et al., 2008; Vorberg and Mattler, 2003). The present results on the ERN speak in favor of a non-linear transition between subjectively *seen* and *unseen* trials: while subliminal performance in both first- and second-order tasks increased smoothly with the target-mask delay (SOA), the ERN did not vary continuously with SOA. Instead, it jumped suddenly as a sole function of subjective visibility showing that the error-detection system reflected by the ERN was strongly impeded for subjectively invisible trials. The crossing of the subjective threshold for conscious reportability was accompanied by a step-like improvement in the availability of information and, more crucially, by the sudden emergence of the ERN. Importantly, the ERN strictly followed the subjective reports of visibility, above and beyond objective variation in stimulation.

These results were obtained by asking participants to subjectively label the trial into two categories, "*seen*" and "*unseen*". This binary visibility judgment was motivated by previous reports showing that in masking paradigms, participants focus their responses on the extreme points of a continuous scale when they are asked to report prime visibility (Sergent and Dehaene, 2004a). Our approach was also adopted for simplicity. Participants already performed no less than three responses on each trial. Requiring them to perform a more complicated visibility rating task would have lengthened the experiment even further. In the future, it might useful to examine if the present findings replicate with a more continuous estimate of visibility (Overgaard et al., 2006; Sergent and Dehaene, 2004a; Sergent et al., 2005; Seth and Dienes, 2008), thus improving our ability to detect whether the ERN follow an all-or-none pattern.

One may raise the critique that subjective reports of visibility are potentially biased and do not accurately reflect the conscious content of the subjects (Persaud et al., 2007). While the issue of finding an appropriate measure of perceptual consciousness remains debated (Lau, 2008; Overgaard et al., 2010; Persaud et al., 2007; Seth et al., 2006) and is not the subject of this study, our results argue that subjective reports provide valid data inasmuch as they correlate strongly with objective changes in behavior and brain activity. Confirming previous results (Del Cul et al., 2007, 2009), we found that visibility reports present a tight correlation with objective performance in the number-comparison task, suggesting that participants are accurately able to monitor and report the state of their perception. Furthermore, our results suggest that subjective reports of visibility reliably index a large objective change in brain activity, namely the ERN. Even when

considering only near-threshold stimuli (intermediate SOA), the ERN switched on or off in tight correlation with subjective reports of visibility or invisibility.

Our results probably go beyond what could have been found using objective measures of visibility alone. Our shortest SOA conditions correspond to fully subliminal trials (Dehaene et al., 2006), since both objective detection and task d' are indistinguishable from zero. We found that these trials are characterized by an absence of ERN and a lack of metacognitive ability. As interesting as such a result might be, it may not be unexpected, considering how much the available sensory evidence is reduced on such heavily masked trials. To determine whether the ERN can be deployed non-consciously, it is therefore crucial to focus on more lightly masked trials, where a longer SOA provides greater sensory evidence for error detection. Unfortunately, such trials provide a challenge for purely objective approaches to consciousness, as their detection d-prime is way above chance. Nevertheless, by sorting trials as a function of whether they fall above or below the threshold for conscious perception, a purely subjective criterion, we found that *unseen* trials are also characterized by an absence of ERN, while at the same time subjects remain better than chance in the metacognitive task of detecting their errors. Interestingly, we show here a complete dissociation between the continuously increasing estimation of error likelihood on *unseen* trials, and the all-or-none detection of errors reflected by the ERN on subjectively *seen* trials.

Computational models of the ERN

How do the brain generators of the ERN compute whether the response is correct or erroneous or a given trial in the absence of any experimenter feedback? Some models of the ERN postulate that it reflects a comparison (Bernstein et al., 1995; Falkenstein et al., 2000) or conflict (Veen and Carter, 2002; Yeung et al., 2004) between the actual and the intended response. How can one integrate awareness in such models? The dual-route model proposed by Del Cul et al. (2009) provides a model of how conscious and non-conscious decisions are made, and how they might be compared to yield an error signal. According to this model, two parallel routes accumulate sensory evidence towards a categorical decision on the same input stimulus. Each route has different noise levels and thresholds: One is a fast, non-conscious sensori-motor route, and one is a slower conscious decision route. A motor response is emitted by the route that first reaches its decision threshold. In the case where time–pressure in emphasized over accuracy, the response is emitted mainly via the fast and noisy motor route which is subject to non-conscious influences (Dehaene et al., 1998; Vorberg and Mattler, 2003). On such trials, the "conscious route" slowly computes the intended response (Del Cul et al., 2009). Any discrepancy between these two responses would then result in an ERN — a difference between intended and executed action. By its very nature, the model generates an ERN only when a conscious intention exists, i.e. when the second route has crossed its threshold. Thus, the model can explain the correlation between conscious perception and the presence of the ERN.

This model is compatible both with the view of the ERN as a conflict monitoring system (Veen and Carter, 2002; Yeung et al., 2004) or a comparison process (Bernstein et al., 1995; Falkenstein et al., 2000). In a similar vein, others have proposed that the ERN is a "prediction-error" signal that indexes the difference between a prediction and an observed outcome: either an ongoing response that departs from the one intended given the perceived stimulus (Alexander and Brown, 2011), or an anticipated reward that departs from the usual one expected when the response is correct (Holroyd and Coles, 2002). Assuming that such expectations are derived from a conscious-level representation of the correct intended response, these mechanisms explain why the ERN is seen only when the stimulus is consciously perceived. On *unseen* trials, no conscious intention or expectation can be computed. Accordingly, the difference process putatively indexed by the ERN is impeded, and cannot distinguish between correct and erroneous responses.

These models also predict that the ERN should vary with the amount of evidence in favor of the correct response and the confidence in the correctness of that response. Indeed, several studies demonstrated a tight correlation between subjective ratings of confidence in one's response, and the size of the ERN (Scheffers and Coles, 2000; Shalgi and Deouell, 2012; Wessel et al., 2011). Scheffers and Coles (2000) showed that for errors due to data limitation, the amplitude of the ERN was identical on correct and error trials. Even within objectively correct responses, the ERN varied massively as a function of whether subjects *believed* that they made an error. Similarly, Shalgi and Deouell (2012) found that for objective errors for which participants were highly confident in their performance rating, the ERN amplitude was predictive of whether the participant thought he had made an error or not. In particular, the ERN vanished when the participant thought he responded correctly, even though the objective performance did not change.

Apparently contradicting the finding, other studies found that it was only a later event-related potential, the Pe, which showed a systematic trial-by-trial correlation with confidence and error awareness. (Dhar et al., 2011; Hughes and Yeung, 2011; O'Connell et al., 2007). Steinhauser and Yeung (2010) demonstrated that financial rewards could shift the participants' threshold for reporting having made an error or a correct response, but that this criterion shift had no impact on the ERN itself. Hughes and Yeung (2011) also found that, while the ERN was reduced in masking conditions, the Pe was the most predictive component of error awareness. In both cases, the ERN remained invariant to changes in error awareness or in error signaling.

Taken together these findings suggest an interesting dissociation between these two components in the global system of performance monitoring. While the ERN seems to reflect a comparison or difference of intended and executed actions (Carbonnell and Falkenstein, 2006) and thus, as we suggest here, varies continuously as a function of intention strength, the Pe seems to be directly linked to the awareness of making an error (Hughes and Yeung, 2011; Nieuwenhuis et al., 2001) and its subsequent signalling (Steinhauser and Yeung, 2010). Such a model predicts that both ERN and Pe should be affected when manipulating the amount of evidence concerning the correct response (Hughes and Yeung, 2011; Maier et al., 2008; Scheffers and Coles, 2000; Shalgi and Deouell, 2012 but see Steinhauser and Yeung, 2012). However, as found by Steinhauser and Yeung (2010), only the Pe should be changed when considering error awareness and subsequent error reportability (Hughes and Yeung, 2011; Nieuwenhuis et al., 2001; Steinhauser and Yeung, 2010). Further analysis of our data on the Pe time-window tended to confirm this hypothesis. While such a model remains speculative and will require further studies to be validated, the present findings provide converging evidence on the role of the ERN in the hierarchy of processes leading to error detection.

Brain regions involved in error monitoring

What brain mechanisms underlie conscious versus non-conscious metacognitive computations? Our results show that error detection is independent of the ERN on *unseen* trials. In both experiments, no ERN was present on *unseen* trials, even when participants correctly evaluated their own performance. In fact, we observed a double dissociation between the ERN and behavioral error detection: no ERN was observed when meta-performance exceeded performance in non-conscious trials (Experiment 1) while the ERN was present even though meta-performance was aligned on performance in conscious trials (Experiment 2). Source reconstruction of the MEEG signal confirmed that activity in one of the main generators of the ERN, the posterior cingulate

cortex (PCC) (Agam et al., 2011; Dhar et al., 2011; Schie et al., 2004), was drastically reduced in the *unseen* condition.

However, on *unseen* trials, brain activity correlating with performance was observed for some of the MEG sensors. Source analysis revealed that this signal originated from the dorsal anterior cingulate cortex (dACC), a region also known to activate after errors (Debener et al., 2005; Dehaene et al., 1994; Keil et al., 2010). Importantly, this activation was present only when time–pressure was relaxed (Experiment 2) and response-times longer, highlighting its sensitivity to evidence accumulation. Activity in this region might thus convey some non-conscious information on the level of confidence in the current response, possibly explaining the participants' subliminal meta-cognitive ability. Note that this brain signal is short-lived and thus may not be sufficient to fully explain above-chance metacognitive responses occurring several hundreds of milliseconds later. However, this activity might be the input to other brain processes that compute the final judgment of confidence in one's response. Brodmann's area 10 is a plausible candidate, as several imaging studies associate it with confidence judgments (Fleming et al., 2010; Rolls et al., 2010; Yokoyama et al., 2010).

Although dACC has long been proposed to be the sole generator of the ERN (Debener et al., 2005; Dehaene et al., 1994; Emeric et al., 2008), our results are compatible with recent evidence suggesting that PCC might be another plausible source for the ERN (Agam et al., 2011; Munro et al., 2007; Vlamings, 2008). Both PCC and dACC have been shown to be active in several error-processing studies (Fassbender et al., 2004; Wittfoth et al., 2008). However it has been suggested that dACC could not only reflect error detection process but might be related to behavioral adjustment such as error avoidance (Magno et al., 2006), mapping between stimulus and response (Williams et al., 2004) and reward prediction-error (Kennerley et al., 2011). Furthermore, dACC has been shown to be activated on conflict trials independently of objective accuracy (Ullsperger and Von Cramon, 2001). Because functional connectivity analyses show that both PCC and dACC are part of a larger functional network (Agam et al., 2011) and share direct anatomical connections (Vogt et al., 2006), it is therefore likely that these regions are both active when an error is made, as suggested by the present MEEG source modelling of experiment 2. Nonetheless, they might have different roles in performance monitoring. A possible framework to explain our data could be that, while PCC directly detects the commission of an error (Agam et al., 2011; Munro et al., 2007; Vlamings, 2008), dACC integrates this information to implement corrective behavior (Modirrousta and Fellows, 2008) and further monitoring processes. While more studies will be needed to pinpoint the functional architecture of cingulate cortex, the present results suggest an interesting difference in sensitivity to conscious versus non-conscious choices for posterior versus anterior cingulate cortex, in keeping with speculations as to the role of the PCC as a crucial node for conscious awareness (Immordino-Yang et al., 2009; Vogt and Laureys, 2009).

Conclusion

Our study suggests the existence of at least two meta-cognitive systems for performance monitoring. One of them is capable of being deployed non-consciously, but it only provides statistical information on the likelihood of having made an error. The other, associated with the ERN, shows an all-or-none signal specifically on error trials where the target was consciously perceived, making it possible for participants to realize their error. By demonstrating the co-existence of these two mechanisms, we provide new evidence on the global architecture of cognitive control and its link to consciousness.

Acknowledgments

We are grateful to the NeuroSpin infrastructure groups, in particular to the doctors Ghislaine Dehaene-Lambertz, Andreas Kleinschmidt, Caroline Huron, Lucie Hertz-Pannier and the nurses Véronique Joly-Testault and Laurence Laurier, for their support in participant recruitment and testing; Virginie van Wassenhove, Marco Buiatti, Leila Rogeau, Etienne Labyt and all the MEG team for their help on technical difficulties; Lauri Parkkonen, Alexandre Gramfort and François Tadel for their assistance on MEEG analysis and source reconstruction; Aaron Schurger and Christophe Pallier for their advice statistical issues; Moti Salti and Simon van Gaal for useful discussions.

This project was supported by a PhD grant of the Direction Générale de l'Armement (DGA, Didier Bazalgette) and a senior grant of the European Research Council to S.D. (NeuroConsc program), as part of a general research program on functional neuroimaging of the human brain (Denis Le Bihan). The NeuroSpin MEG facility was sponsored by grants from INSERM, CEA, the Fondation pour la Recherche Médicale, the Bettencourt-Schueller foundation, and the Région île-de-France. F.V.O. is a Postdoctoral Fellow of the Research Foundation — Flanders (FWO-Vlaanderen). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.neuroimage.2013.01.054.

References

Agam, Y., Hamalainen, M., Lee, A.C.H., Dyckman, K.A., Friedman, J.S., Isom, M., Makris, N., Manoach, D.S., 2011. Multimodal neuroimaging dissociates hemodynamic and electrophysiological correlates of error processing. Proc. Natl. Acad. Sci. 108, 17556–17561.
Alexander, W.H., Brown, J.W., 2011. Medial prefrontal cortex as an action-outcome predictor. Nat. Neurosci. 14, 1338–1344.
Aly, M., Yonelinas, A.P., 2012. Bridging consciousness and cognition in memory and perception: evidence for both state and strength processes. PLoS One 7, e30231.
Baayen, R.H., Davidson, D.J., Bates, D.M., 2008. Mixed-effects modeling with crossed random effects for subjects and items. J. Mem. Lang. 59, 390–412.
Bernstein, P.S., Scheffers, M.K., Coles, M.G.H., 1995. "Where did I go wrong?" A psychophysiological analysis of error detection. J. Exp. Psychol. Hum. Percept. Perform. 21, 1312–1322.
Botvinick, M.M., Braver, T.S., Barch, D.M., Carter, C.S., Cohen, J.D., 2001. Conflict monitoring and cognitive control. Psychol. Rev. 108, 624–652.
Carbonnell, L., Falkenstein, M., 2006. Does the error negativity reflect the degree of response conflict? Brain Res. 1095, 124–130.
Debener, S., Ullsperger, M., Siegel, M., Fiehler, K., Cramon, D.Y. Von, Engel, A.K., Von Cramon, D.Y., 2005. Trial-by-trial coupling of concurrent electroencephalogram and functional magnetic resonance imaging identifies the dynamics of performance monitoring. J. Neurosci. 25, 11730–11737.
Dehaene, S., Changeux, J.-P., 2011. Experimental and theoretical approaches to conscious processing. Neuron 70, 200–227.
Dehaene, S., Posner, M.I., Tucker, D.M., 1994. Localization of a neural system for error detection and compensation. Psychol. Sci. 5, 303–305.
Dehaene, S., Naccache, L., Le Clec'H, G., Koechlin, E., Mueller, M., Dehaene-Lambertz, G., Van de Moortele, P.F., Le Bihan, D., 1998. Imaging unconscious semantic priming. Nature 395, 597–600.
Dehaene, S., Naccache, L., Cohen, L., Bihan, D. Le, Mangin, J.F., Poline, J.-B., Riviere, D., 2001. Cerebral mechanisms of word masking and unconscious repetition priming. Nat. Neurosci. 4, 752–758.
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.
Del Cul, A., Baillet, S., Dehaene, S., 2007. Brain dynamics underlying the nonlinear threshold for access to consciousness. PLoS Biol. 5, 2408–2423.
Del Cul, A., Dehaene, S., Reyes, P., Bravo, E., Slachevsky, A., 2009. Causal role of prefrontal cortex in the threshold for access to consciousness. Brain 132, 2531–2540.
Dhar, M., Wiersema, J.R., Pourtois, G., 2011. Cascade of neural events leading from error commission to subsequent awareness revealed using EEG source imaging. PLoS One 6, e19578.
Emeric, E.E., Brown, J.W., Leslie, M., Pouget, P., Stuphorn, V., Schall, J.D., 2008. Performance monitoring local field potentials in the medial frontal cortex of primates: anterior cingulate cortex. J. Neurophysiol. 99, 759–772.
Endrass, T., Reuter, B., Kathmann, N., 2007. ERP correlates of conscious error recognition: aware and unaware errors in an antisaccade task. Eur. J. Neurosci. 26, 1714–1720.
Evans, S., Azzopardi, P., 2007. Evaluation of a 'bias-free' measure of awareness. Spat. Vis. 20 (20), 61–77.
Falkenstein, M., Hoormann, J., Christ, S., Hohnsbein, J., 2000. ERP components on reaction errors and their functional significance: a tutorial. Biol. Psychol. 51, 87–107.
Fassbender, C., Murphy, K., Foxe, J.J., Wylie, G.R., Javitt, D.C., Robertson, I.H., Garavan, H., 2004. A topography of executive functions and their interactions revealed by functional magnetic resonance imaging. Brain Res. Cogn. Brain Res. 20, 132–143.

Fleming, S.M., Dolan, R.J., 2010. Effects of loss aversion on post-decision wagering: implications for measures of awareness. Conscious. Cogn. 19, 352–363.
Fleming, S.M., Weil, R.S., Nagy, Z., Dolan, R.J., Rees, G., 2010. Relating introspective accuracy to individual differences in brain structure. Science 329, 1541–1543.
Galvin, S.J., Podd, J.V., Drga, V., Whitmore, J., 2003. Type 2 tasks in the theory of signal detectability: discrimination between correct and incorrect decisions. Psychon. Bull. Rev. 10, 843–876.
Gehring, W.J., Fencsik, D., 2001. Functions of the medial frontal cortex in the processing of conflict and errors. J. Neurosci. 21, 9430.
Gehring, W.J., Goss, B., Coles, M.G.H., Meyer, D.E., Donchin, E., 1993. A neural system for error detection and compensation. Psychol. Sci. 4, 385–390.
Hewig, J., Coles, M., Trippe, R., 2011. Dissociation of Pe and ERN/Ne in the conscious recognition of an error. Psychophysiology 1–7.
Holroyd, C.B., Coles, M.G.H., 2002. The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. Psychol. Rev. 109, 679–709.
Hughes, G., Yeung, N., 2011. Dissociable correlates of response conflict and error awareness in error-related brain activity. Neuropsychologia 49, 405–415.
Immordino-Yang, M.H., McColl, A., Damasio, H., Damasio, A., 2009. Neural correlates of admiration and compassion. Proc. Natl. Acad. Sci. U. S. A. 106, 8021–8026.
Irimia, A., Van Horn, J.D., Halgren, E., 2011. Source cancellation profiles of electroencephalography and magnetoencephalography. NeuroImage 59, 2464–2474.
Kanai, R., Walsh, V., Tseng, C.-H., 2010. Subjective discriminability of invisibility: a framework for distinguishing perceptual and attentional failures of awareness. Conscious. Cogn. 19, 1045–1057.
Keil, J., Weisz, N., Paul-Jordanov, I., Wienbruch, C., 2010. Localization of the magnetic equivalent of the ERN and induced oscillatory brain activity. NeuroImage 51, 404–411.
Kennerley, S.W., Behrens, T.E.J., Wallis, J.D., 2011. Double dissociation of value computations in orbitofrontal and anterior cingulate neurons. Nat. Neurosci. 14, 1581–1589.
Kentridge, R., Heywood, C., 1999. The status of blindsight: near-threshold vision, islands of cortex and the Riddoch phenomenon. J. Conscious. Stud. 6, 3–11.
Kepecs, A., Uchida, N., Zariwala, H.A., Mainen, Z.F., 2008. Neural correlates, computation and behavioural impact of decision confidence. Nature 455, 227–231.
Kiani, R., Shadlen, M.N., 2009. Representation of confidence associated with a decision by neurons in the parietal cortex. Science 324, 759–764.
Kouider, S., Dehaene, S., 2007. Levels of processing during non-conscious perception: a critical review of visual masking. Philos. Trans. R. Soc. Lond. B: Biol. Sci. 362, 857–875.
Kunimoto, C., Miller, J., Pashler, H.E., 2001. Confidence and accuracy of near-threshold discrimination responses. Conscious. Cogn. 10, 294–340.
Lau, H.C., 2008. A higher order Bayesian decision theory of consciousness. Prog. Brain Res. 168, 35–48.
Lau, H.C., Passingham, R.E., 2006. Relative blindsight in normal observers and the neural correlate of visual consciousness. Proc. Natl. Acad. Sci. U. S. A. 103, 18763–18768.
Lau, H.C., Passingham, R.E., 2007. Unconscious activation of the cognitive control system in the human prefrontal cortex. J. Neurosci. 27, 5805–5811.
Lau, H.C., Rosenthal, D., 2011. Empirical support for higher-order theories of conscious awareness. Trends Cogn. Sci. 15, 365–373.
Logan, G.D., Crump, M.J.C., 2010. Cognitive illusions of authorship reveal hierarchical error detection in skilled typists. Science 330, 683.
Magno, E., Foxe, J.J., Molholm, S., Robertson, I.H., Garavan, H., 2006. The anterior cingulate and error avoidance. J. Neurosci. 26, 4769–4773.
Maier, M.E., Steinhauser, M., Hubner, R., 2008. Is the error-related negativity amplitude related to error detectability? Evidence from effects of different error types. J. Cogn. Neurosci. 20, 2263–2273.
Maniscalco, B., Lau, H.C., 2012. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. Conscious. Cogn. 21, 422–430.
Maris, E., Oostenveld, R., 2007. Nonparametric statistical testing of EEG- and MEG-data. J. Neurosci. Methods 164, 177–190.
Melloni, L., Molina, C., Pena, M., Torres, D., Singer, W., Rodriguez, E., 2007. Synchronization of neural activity across cortical areas correlates with conscious perception. J. Neurosci. 27, 2858–2865.
Miltner, W.H., Lemke, U., Weiss, T., Holroyd, C.B., Scheffers, M.K., Coles, M.G.H., 2003. Implementation of error-processing in the human anterior cingulate cortex: a source analysis of the magnetic equivalent of the error-related negativity. Biol. Psychol. 64, 157–166.
Modirrousta, M., Fellows, L.K., 2008. Dorsal medial prefrontal cortex plays a necessary role in rapid error prediction in humans. J. Neurosci. 28, 14000–14005.
Munro, G.E.S., Dywan, J., Harris, G.T., McKee, S., Unsal, A., Segalowitz, S.J., 2007. ERN varies with degree of psychopathy in an emotion discrimination task. Biol. Psychol. 76, 31–42.
Nieuwenhuis, S., Ridderinkhof, K.R., Blom, J.H., Band, G.P.H., Kok, A., 2001. Error-related brain potentials are differentially related to awareness of response errors: evidence from an antisaccade task. Psychophysiology 38, 752–760.
Nieuwenhuis, S., Schweizer, T.S., Mars, R.B., Botvinick, M.M., Hajcak, G., 2007. Error-likelihood prediction in the medial frontal cortex: a critical evaluation. Cereb. Cortex 17, 1570–1581.
O'Connell, R.G., Dockree, P.M., Bellgrove, M.A., Kelly, S.P., Hester, R., Garavan, H., Robertson, I.H., Foxe, J.J., 2007. The role of cingulate cortex in the detection of errors with and without awareness: a high-density electrical mapping study. Eur. J. Neurosci. 25, 2571–2579.
Overgaard, M., Rote, J., Mouridsen, K., Ramsøy, T.Z., 2006. Is conscious perception gradual or dichotomous? A comparison of report methodologies during a visual task. Conscious. Cogn. 15, 700–708.
Overgaard, M., Timmermans, B., Sandberg, K., Cleeremans, A., 2010. Optimizing subjective measures of consciousness. Conscious. Cogn. 19, 682–686.
Pavone, E.F.E.F., Marzi, C.A., Girelli, M., 2009. Does subliminal visual perception have an error-monitoring system? Eur. J. Neurosci. 30, 1424–1431.
Persaud, N., McLeod, P., Cowey, A., 2007. Post-decision wagering objectively measures awareness. Nat. Neurosci. 10, 257–261.
Pessiglione, M., Schmidt, L., Draganski, B., Kalisch, R., Lau, H.C., Dolan, R.J., Frith, C.D., 2007. How the brain translates money into force: a neuroimaging study of subliminal motivation. Science 316, 904.
Pleskac, T.J., Busemeyer, J.R., 2010. Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. Psychol. Rev. 117, 864–901.
Province, J.M., Rouder, J.N., 2012. Evidence for discrete-state processing in recognition memory. Proc. Natl. Acad. Sci. U. S. A. 109, 14357–14362.
Quiroga, R.Q., Mukamel, R., Isham, E.A., Malach, R., Fried, I., 2008. Human single-neuron responses at the threshold of conscious recognition. Proc. Natl. Acad. Sci. U. S. A. 105, 3599–3604.
Resulaj, A., Kiani, R., Wolpert, D.M., Shadlen, M.N., 2009. Changes of mind in decision-making. Nature 461, 263–266.
Rolls, E.T., Grabenhorst, F., Deco, G., 2010. Choice, difficulty, and confidence in the brain. NeuroImage 53, 694–706.
Rounis, E., Maniscalco, B., Rothwell, J.C., Passingham, R.E., Lau, H.C., 2010. Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. Cogn. Neurosci. 1, 165–175.
Scheffers, M.K., Coles, M.G.H., 2000. Performance monitoring in a confusing world: error-related brain activity, judgments of response accuracy, and types of errors. J. Exp. Psychol. Hum. Percept. Perform. 26, 141–151.
Schie, H.T. Van, Mars, R.B., Coles, M.G.H., Bekkering, H., Van Schie, H.T., 2004. Modulation of activity in medial frontal and motor cortices during error observation. Nat. Neurosci. 7, 549–554.
Sergent, C., Dehaene, S., 2004a. Is consciousness a gradual phenomenon? Evidence for an all-or-none bifurcation during the attentional blink. Psychol. Sci. 15, 720–728.
Sergent, C., Dehaene, S., 2004b. Neural processes underlying conscious perception: experimental findings and a global neuronal workspace framework. J. Physiol. Paris 98, 374–384.
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.
Seth, A.K., Dienes, Z., 2008. Measuring consciousness: relating behavioural and neurophysiological approaches. Trends Cogn. Sci. 12, 314–321.
Seth, A.K., Izhikevich, E., Reeke, G.N., Edelman, G.M., 2006. Theories and measures of consciousness: an extended framework. Proc. Natl. Acad. Sci. U. S. A. 103, 10799–10804.
Shalgi, S., Deouell, L.Y., 2012. Is any awareness necessary for an Ne? Front. Hum. Neurosci. 6, 1–15.
Steinhauser, M., Yeung, N., 2010. Decision processes in human performance monitoring. J. Neurosci. 30, 15643–15653.
Steinhauser, M., Yeung, N., 2012. Error awareness as evidence accumulation: effects of speed-accuracy trade-off on error signaling. Front. Hum. Neurosci. 6, 240.
Ullsperger, M., Von Cramon, D.Y., 2001. Subprocesses of performance monitoring: a dissociation of error processing and response competition revealed by event-related fMRI and ERPs. NeuroImage 14, 1387–1401.
Van den Bussche, E., Notebaert, K., Reynvoet, B., 2009. Masked primes can be genuinely semantically processed: a picture prime study. Exp. Psychol. 56, 295–300.
Van Gaal, S., Ridderinkhof, K.R., Fahrenfort, J.J., Scholte, H.S., Lamme, V.A.F., 2008. Frontal cortex mediates unconsciously triggered inhibitory control. J. Neurosci. 28, 8053–8062.
Veen, V. Van, Carter, C.S., 2002. The anterior cingulate as a conflict monitor: fMRI and ERP studies. Physiol. Behav. 77, 477–482.
Vlamings, P., 2008. Reduced error monitoring in children with autism spectrum disorder: an ERP study. Eur. J. Neurosci. 28, 399–406.
Vogt, B.A., Laureys, S., 2009. Posterior cingulate, precuneal & retrosplenial cortices: cytology & components of the neural network correlates of consciousness. Brain 6123, 205–217.
Vogt, B., Vogt, L., Laureys, S., 2006. Cytology and functionally correlated circuits of human posterior cingulate areas. NeuroImage 29, 452–466.
Vorberg, D., Mattler, U., 2003. Different time courses for visual perception and action priming. Proc. Natl. Acad. Sci. 100, 6275–6280.
Weiskrantz, L., 1996. Blindsight revisited. Curr. Opin. Neurobiol. 6, 215–220.
Wessel, J., Danielmeier, C., Ullsperger, M., 2011. Error awareness revisited: accumulation of multimodal evidence from central and autonomic nervous systems. J. Cogn. Neurosci. 23, 3021–3036.
Williams, Z.M., Bush, G., Rauch, S.L., Cosgrove, G.R., Eskandar, E.N., 2004. Human anterior cingulate neurons and the integration of monetary reward with motor responses. Nat. Neurosci. 7, 1370–1375.
Wittfoth, M., Küstermann, E., Fahle, M., Herrmann, M., 2008. The influence of response conflict on error processing: evidence from event-related fMRI. Brain Res. 1194, 118–129.
Woodman, G.F.F., 2010. Masked targets trigger event-related potentials indexing shifts of attention but not error detection. Psychophysiology 47, 410–414.
Yeung, N., Botvinick, M.M., Cohen, J.D., 2004. The neural basis of error detection: conflict monitoring and the error-related negativity. Psychol. Rev. 111, 931–959.
Yokoyama, O., Miura, N., Watanabe, J., Takemoto, A., Uchida, S., Sugiura, M., Horie, K., Sato, S., Kawashima, R., Nakamura, K., 2010. Right frontopolar cortex activity correlates with reliability of retrospective rating of confidence in short-term recognition memory performance. Neurosci. Res. 68, 199–206.

Supplementary Materials for:

Distinct brain mechanisms for conscious versus subliminal error detection

Lucie Charles, Filip van Opstal, Sébastien Marti and Stanislas Dehaene

Content list:

Supplementary Methods

- RT correction method

Supplementary Results

- Additional behavioral analyses
  - Analysis of response time (**Figures S1**)
  - Could RT variations explain meta-performance? (**Figures S2 and S3**)

- Additional MEEG analyses
  - ERP analysis before RT correction (**Figure S4**)
  - Correct meta-performance in the absence of the ERN
  - Brain activity in subliminal meta-correct trials (**Figure S5**)
  - Analysis of the Pe time-window

SUPPLEMENTARY METHODS

RT correction method

To obtain event-related responses with equalized response times on correct and incorrect trials in experiment 1, we used a trial averaging method that weighted identically error and correct trials with similar reaction time, and removed trials whose RT did not match any RT of the opposite category. For each subject, we compiled 20 ms time-bins histograms of RTs, separately for correct (*c*) and error (*e*) trials, and computed for each trial category the mean response-locked MEEG responses $\mu_c(b)$ and $\mu_e(b)$ in each bin *b*, and the corresponding number of trials $n_c(b)$ and $n_e(b)$. We discarded bins where only one category of trial (either correct or error) was observed, i.e. those in which either $n_c(b)=0$ or $n_e(b)=0$. The remaining bins were used to compute an equally weighted mean, using as weight the total number of trials in each bin, i.e. $n_{total}(b) = n_c(b) + n_e(b)$. Thus, for error trials, the evoked response was calculated as explained in equation 1:

[1] $$\mu_e = \frac{\sum_b n_{total}(b)\mu_e(b)}{\sum_b n_{total}(b)}$$

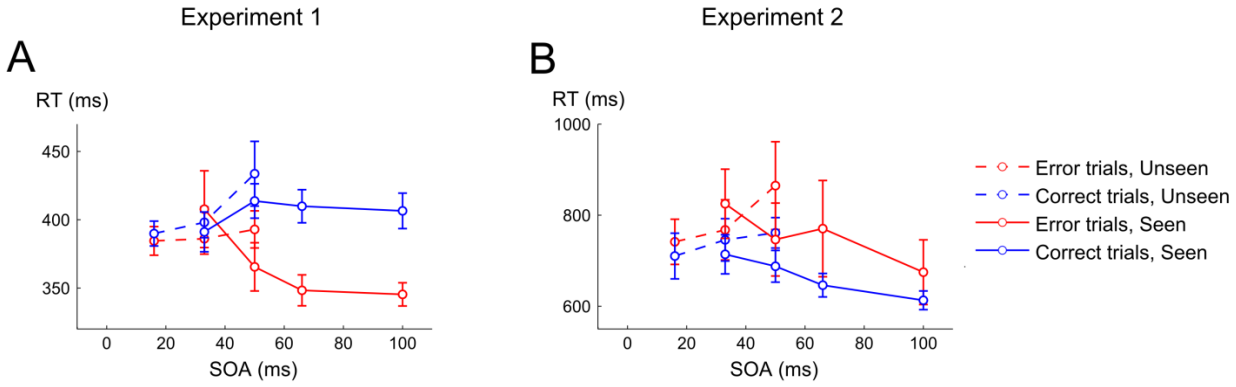
The symmetrical equation, switching *e* and *c* indices, was applied for correct trials.

SUPPLEMENTARY RESULTS

Additional behavioral analyses

Analysis of response times

Supplementary Figure 1. Response-times from experiment 1 (left column) and 2 (right column). (A-B) Median reaction times were computed for error (red lines) and correct (blue lines) trials, separately for *seen* (solid lines) and *unseen* trials. Data points with insufficient numbers of measures were excluded (see Methods). Error bars represent standard-error.



Median RTs were submitted to a linear mixed-effects model (see Methods) with SOA (5 levels: 16, 33, 50, 66 and 100 ms), visibility (*seen* or *unseen*) and performance (*correct* or *error*) as factors. A significant main effect of visibility was found in experiment 1 ($p = 0.018$) and in experiment 2 ($p=0.033$) as RTs were overall shorter for seen than for unseen trials in both experiments (median RTs of 383 ms vs 402 ms for experiment 1 and 740 ms vs 818 ms in experiment 2). The main effect of performance only approached significance in experiment 1 ($p= 0.0672$), error trials corresponding overall to shorter RTs (median : 365 ms) than correct trials (median: 406 ms), while a trend in the opposite direction was observed in experiment 2 (slow error trials with median RTs of 747 ms and fast correct trials with median RTs of 696 ms) but did not reach significance ($p=0.87$). The main effect of SOA did not reach significance in any of the experiments ($p = 0.76$ and $p = 0.09$, respectively). However significant interactions between visibility and SOA were found in both experiments ($p=0.001$

and p=0.01), as RTs tended to increase with SOA on unseen trials (p=0.06 in exp. 1; n.s. in exp. 2) and to decrease with SOA on seen trials (n.s. in exp. 1; p=0.004 in exp. 2).

The interaction between performance and SOA reached significance for experiment 1 (p=0.005), corresponding to the fact that error RTs decreased with SOA ($F_{4,70}$ = 2.20, p=0.077, near significance) while correct RTs did not, a result absent for experiment 2 (p=0.82).

No interaction between performance and visibility was found in any of the two experiments. Reducing the analysis to *unseen* trials, error trials were significantly faster than correct trials in experiment 1 (t_{12}=-2.28, p=0.042) while a trend in the opposite direction did not reach significance in experiment 2 (t_{12}=1.44, p=0.17). Similarly for s*een* trials, error trials were significantly faster than correct trials in experiment 1 (t_{12}=-4.64, p=0.0005) while the opposite effect was found in experiment 2 (t_{12}=2.44, p=0.031). Overall, this pattern is consistent with the different time-pressure instructions given in each experiment: fast errors were obtained under strong time pressure in experiment 1, while errors were associated with slow RTs when time pressure was relaxed in experiment 2.

Could RT variations explain meta-performance?

Since RTs were significantly faster on error than on correct trials in the unseen trials of experiment 1, we wondered whether this factor alone could explain the above-chance meta-performance in unconscious error monitoring. Perhaps subjects simply monitored their own RT on individual trials, and used it as an indicator of their accuracy.

Note first that this interpretation is unlikely as a global interpretation of our results because, in experiment 2, error trials were (non-significantly) slower than correct trials, and yet meta-performance still remained above chance. Furthermore, at some SOAs, RT differences were arguably too small or inexistent to support the observed meta-performance

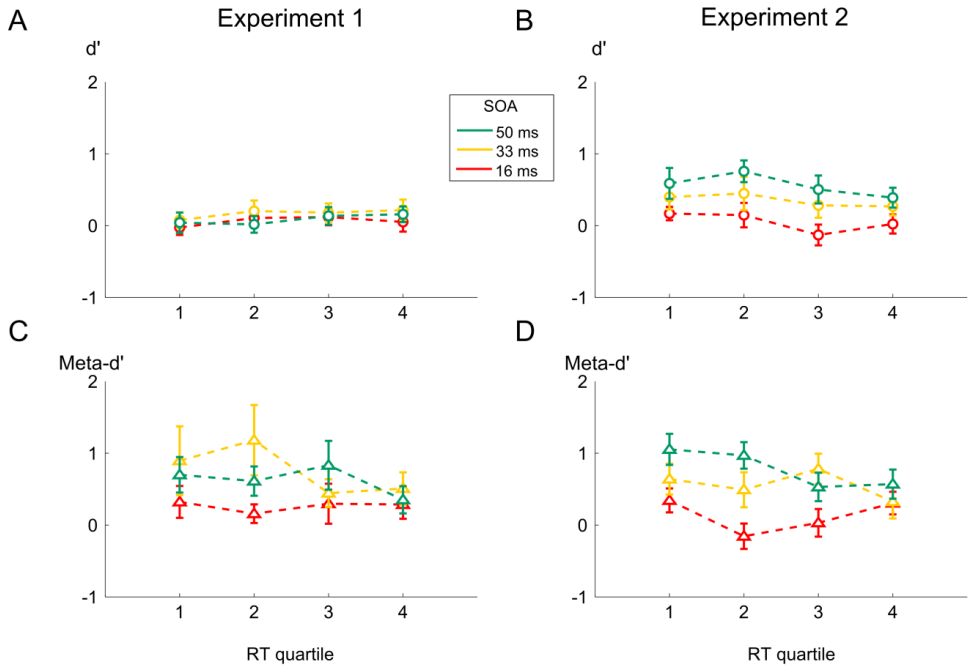
(See figure S1A-B). In experiment 1, there was a genuine parallel between RT and meta-performance, as RTs were not significantly different for error and correct trials at SOA=16 ms (t_{12}= -0.93, p=0.37), where meta-performance was at chance, while the RT difference approached significance for SOA 33 ms (t_{12}= -2.14, p = 0.05) and SOA 50 ms (t_{11}= -1.94, p = 0.08) where meta-performance was significant. However, in experiment 2, although meta-performance was consistently above chance at SOA 33 and 50 ms, no significant difference between error and correct conditions was found in any SOA condition (t_{12}= 1.56, p=0.14 for SOA 16 ms, t_{12}= 0.7, p = 0.28 for SOA 33 ms and t_{12}= 1.12, p = 0.28 for SOA 50 ms). These findings make it unlikely that subjects used RT to predict their performance.

To obtain more decisive evidence on this point, we performed two additional analyses which estimated whether meta-performance remained above chance when RT could not be used to predict accuracy. First, for unseen trials only, we separated the RT distribution of each SOA and each subject into four quartiles. We then computed d' and meta-d' separately for trials within each such quartile, focusing on the two intermediates quartile (intervals 25-50% and 50-75%) where RT variation was minimized (Figure S2). In these two quartiles, if subjects used RTs to predict their performance, their meta-performance should be at chance. At the shortest SOA (16 ms), meta-d' values were indeed not significantly different from 0 for any quartile in both experiments. However, meta-d' was significantly different from 0 for the two intermediates quartile for SOA 33 ms for experiment 1 (t_{12}=2.41, p=0.03 and t_{12}=2.29,p=0.04) and experiment 2 (t_{12}=2.01, p=0.067 and t_{12}=3.7, p=0.003). Similar above-chance meta-performance was observed for SOA 50 ms in experiment 1 (t_{12}=3.0, p=0.01 and t_{12}=2.45, p=0.03) and in experiment 2 (t_{12}=5.25, p=0.0002 and t_{12}=2.6847, p=0.02).

Additionally, we performed an ANOVA on meta-performance with quartiles and SOA as factors, to study a possible interaction between SOA and quartile. In experiment 1, this analysis revealed no main effect of SOA ($F_{2,24}$= 1.71, p=0.20), no main effect of quartile

($F_{3,36}$=0.83, p=0.48) and no interaction between the two ($F_{6,72}$=1.10, p=0.37). Indeed, no clear pattern for the effect of SOA could be observed in the data (Figure S2). For experiment 2, we found a main effect of SOA ($F_{2,24}$=5.76, p=0.009), showing that longer SOA were associated with better meta-performance. However, the analysis revealed no main effect of quartile ($F_{2,24}$=2.207, p=0.10) and only a near-threshold interaction between SOA and quartile ($F_{3,36}$=2.10, p=0.063). Overall, these results suggested that the quartile did not introduce any difference in meta-performance in any of the experiment, making it unlikely that the subject used their RTs to predict their performance.

Supplementary Figure 2 : d' and meta-d' as a function of RT quartile. Unbiased measures of performance (d', circles, top raw) and meta-performance (meta-d', triangles, bottom raw) as a function of RT quartile for each condition of SOA (50 ms green line, 33 ms yellow line and 16 ms red line) of *unseen* trials, from experiment 1 (left column) and 2 (right column).



As a second, more stringent control, we sorted the unseen trials as a function of whether the RT above or below the median RT for this SOA. We then systematically crossed

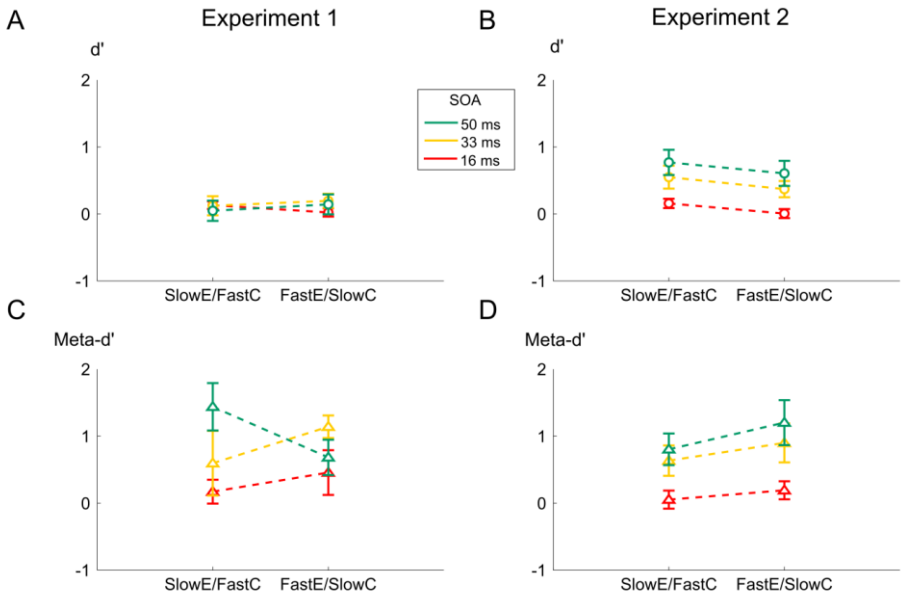
performance and RT by creating two sets of trials, one in which errors were slow and correct trials were fast, and the other in which the converse was true. If subjects relied solely on a strategy of monitoring their RT to detect their subliminal errors, then meta-performance should strongly vary across these two sets and should drop *below* chance in one of these two sets of trials (the one for which error and correct RTs were opposite to their strategy). This effect, however, was not observed.

In experiment 1, meta-performance did not differ significantly on SlowError/FastCorrect and FastError/SlowCorrect data sets, for SOA 16 ms (t_{12}=-0.73, p=0.48) and SOA 33 ms (t_{12}=-1.41, p=0.18). For SOA 50 ms, a significant difference was observed (t_{12}=2.92, p=0.012), but in the direction opposite to that expected from the RT model: meta-performance was actually better in the SlowError/FastCorrect condition, although errors were overall faster than correct responses in the "seen" condition and thus this condition should have misled subjects into thinking that their slow responses were correct. Note that, contrary to the predictions of the RT model, meta-performance never fell below chance. Quite the contrary, in the crucial SlowError/FastCorrect data set, meta-d' was significantly *above* chance from 0 for SOA 50 ms (t_{12}=4.06, p=0.001), although the effect did not reach significance for SOA 16 ms (t_{12}=0.96, p=0.35) nor SOA 33 ms (t_{12}=1.23, p=0.24). An ANOVA on meta-d' with SOA and RT condition as factors showed no significant effect of RT condition ($F_{1,12}$=0.012, p=0.91), even when focusing only on the 16 and 33 ms SOAs ($F_{1,12}$=1.9, p=0.19), thus providing no evidence that the RT strategy was used.

In experiment 2, again, meta-performance was not significantly different in the FastError/SlowCorrect than in SlowError/FastCorrect condition for any of the SOA condition (SOA 16 ms: t_{12}=-0.57, p=0.58; SOA 33 ms : t_{12}=-0.95, p=0.36; SOA 50 ms : t_{12}=-1.39, p=0.19). Errors were overall slower than correct trials in the "seen" condition, and thus the RT strategy should have predicted below-chance meta-performance in the

FastError/SlowCorrect data set. However, this prediction was systematically violated, as meta-d' was significantly *above* chance for SOA 33 ms (t_{12}=3.09, p=0.009) and SOA 50 ms (t_{12}=3.59, p=0.004), though not for SOA 16 ms (t_{12}=0.39, p=0. 0.70). An ANOVA on meta-d' with SOA and RT as factors revealed no main effect of RT condition ($F_{1,12}$=2.23, p=0.16). In brief, the pattern of results makes it highly unlikely that subjects used a strategy of monitoring their own RTs to evaluate the accuracy of their response.

Supplementary Figure 3 : Unbiased measures of performance (d', circles, top raw) and meta-performance (meta-d', triangles, bottom raw) as a function of RT condition. Points on the left represent conditions where errors trial had RTs above the median and correct trials had RTs below the median of the RT distribution while points on the right represent the opposite pattern. Results are displayed for each condition of SOA (50 ms green line, 33 ms yellow line and 16 ms red line) of *unseen* trials, for experiment 1 (left column) and 2 (right column).



Additional MEEG analyses

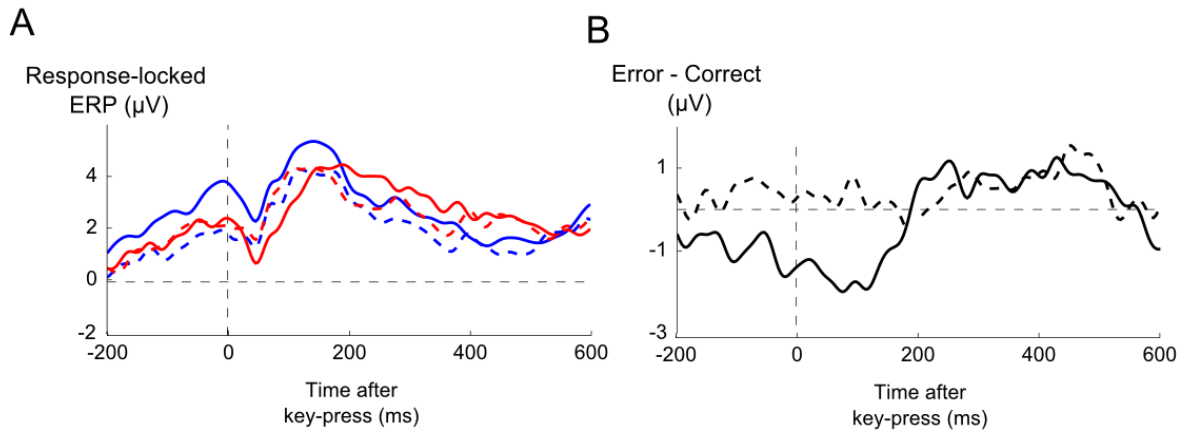
ERP Analysis before RT correction

To ensure that the RT correction method used in experiment 1 did not modify qualitatively the results, we performed an identical analysis on uncorrected ERPs. Supplementary Figure 3 shows the uncorrected response-locked ERPs on a fronto-central cluster of electrodes for seen and unseen trials. One can observe a shift in baseline for seen correct trials, probably due to the superimposition of other sensory-evoked ERP components on response-related signal. Indeed, the ramping effect observed for all conditions suggests that the motor response occurred while stimulus-evoked components were still present. A shift of these ramping components, because RT was shorter on error than on correct trials, can explain the observed baseline shift. Note that this effect had a smaller impact on the unseen condition, where RTs were overall more similar for error and correct trials. Comparing figure 4 with figure S3, one also sees that our RT correction procedure was successful in removing these baseline deviations.

Crucially, these artifactual baseline shifts did not alter our main observations concerning the ERN. In *seen* trials, the ERN was clearly present as a more negative voltage on error than on correct trials (t_{12}=-3.73, p=0.002). Furthermore, no significant difference was detectable in *unseen* trials (t_{12}=0.70, p=0.49), confirming our RT-corrected analyses and resulting in a significant interaction between visibility (*seen* or *unseen*) and performance (error or correct trials) ($F_{1,36}$=9,39, p=0.009). These findings indicate that the RT correction method did not qualitatively modify the results, and allow us to affirm that our results cannot be attributed to RT differences or to the baseline differences that they may cause.

Supplementary Figure 4. Uncorrected time course of event-related potentials as a function of objective performance and subjective visibility. (A) Grand-average event-related potentials

(ERPs) recorded from a cluster of central electrodes (FC1, FC2, C1, Cz, C2). (B) Difference waveforms of error minus correct trials, separately for *seen* (solid line) and *unseen* (dashed line) trials.



Correct meta-performance in the absence of the ERN

Behavioral analysis indicated that, even on *unseen* trials, participants were above chance in performance evaluation. One possibility is that an ERN might indeed be present in unconscious condition but only for trials with good metacognitive performance. Therefore, we analysed more closely the trials classified as unseen but with correct meta-cognitive judgments (i.e. trials considered as hits in the second-order decision).

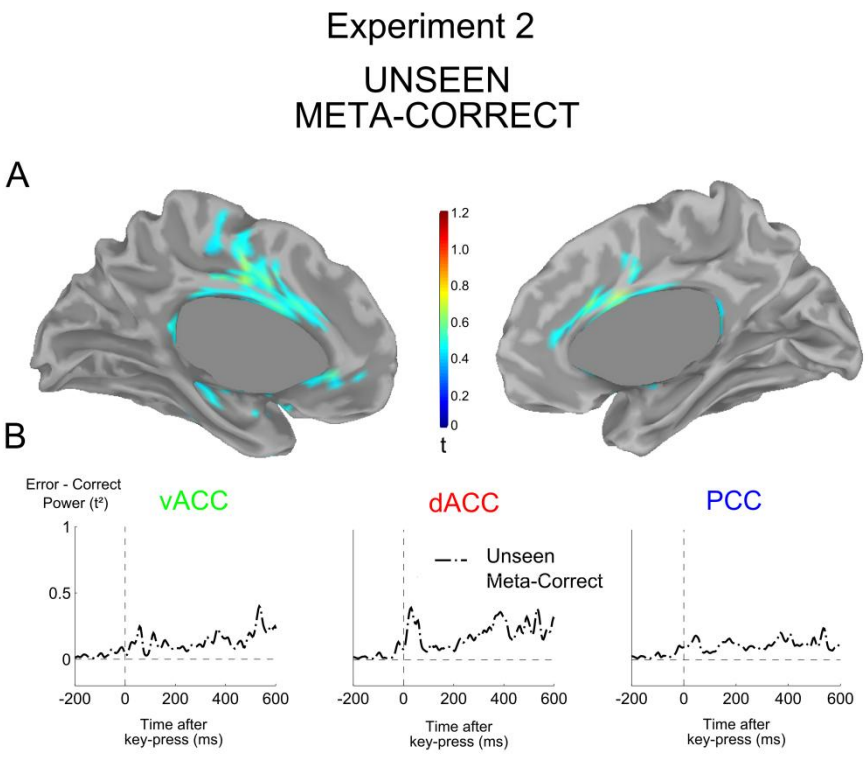
As previously, we first looked at the a-priori cluster of central electrodes (FC1, FC2, C1, Cz, C2) and over the 0-150 ms after motor response time-window. No significant difference between error and correct was found for *unseen* trials in both experiments (all $p > 0.30$). We also searched for significant clusters discriminating between error and correct trials analysis on both MEG and EEG sensors in the same time window. In experiment 1, no cluster was found in any type of sensor. In experiment 2, two significant clusters similar to the ones found previously in the overall *unseen* condition were found in MEG data: a left central cluster for magnetometers ($p=0.014$), and an occipito-central cluster in longitudinal

gradiometers (p=0.016). However, this difference in activation appeared on a slightly earlier time-window (from 0 to 50 ms and 0 to 70 ms) than the classic peak of the ERN. Therefore, these results suggest that above-chance meta-cognitive performance did not rely on the ERN.

Brain activity in subliminal meta-correct trials

In order to further investigate what might be the neural substrate of the above-chance meta-performance in subliminal trials, we estimated the sources of the MEEG signal in the cortex specifically for trials where performance was correctly evaluated. To do so, within the subliminal trials, we studied the difference in brain activity between error trials classified as errors and correct trials classified as correct. In experiment 1, no significant pattern of activity could be found. In experiment 2, activity was found in dorsal anterior cingulate gyrus (Brodman area 24, Supplementary Figure S4A) peaking 30 ms after the button press (talairach coordinates x= 6 y=7 z=28), with an amplitude of approximately 40% of the peak observed for seen trials. Increased activity after the response was also found in PCC and vACC, yet without peaking synchronously to the ERN.

Supplementary Figure 5: Difference of source estimates between errors classified as errors and correct trials classified as correct, on subliminal trials in experiment 2 (i.e. trials with accurate subliminal meta-performance). (A) View of the medial surface of the left and right hemispheres. Data are thresholded at 66% of maximum activity of each condition. Brain activity was average on the 0-100 ms time-window. (B) Time course of brain activity in three bilateral regions of interest located in ventral Anterior Cingulate Cortex (vACC), dorsal Anterior Cingulate Cortex (dACC) and Posterior Cingulate Cortex (PCC), line). Values correspond to instantaneous power in the region of interest (ie average across vertices of the square of brain signal).



Analysis of the Pe time-window

In order to examine more closely further differences between error and correct trials, we then performed an analysis on a time-window corresponding to the Pe (Pe=positivity on errors), on an a priori cluster of electrodes (FC1, FC2, C1, Cz, C2). In experiment 1, no significant difference was found, neither for the *seen* nor the *unseen* trials. Surprisingly, this result remained identical even when considering only meta-correct trials, where subject accurately reported their performance. Nevertheless, a significant difference was found for experiment 2 on a time-window of 200-400 ms after the motor response both for *seen* ($p = 0.013$) and *unseen* ($p = 0.026$) conditions. As the Pe is often associated with the awareness of making an error, this activity might thus be one of the correlate of above-chance estimation in *unseen* trials.

While no Pe was observed in experiment 1 neither in the *seen* nor in the *unseen* condition, we investigated more closely if this result was still true when splitting the

conditions by SOA. We focused on the narrower time-window of 200-300 ms and we observed that for *seen* trials, after making an error the positivity was greater for largest SOAs than for shorter SOAs. These results were confirmed by an ANOVA where we observed a main effect of SOA for *seen* trials ($p<10^{-3}$) and a near-significance interaction between SOA and performance (p=0.051). Indeed the difference between error and correct trials almost reached significance for SOA 66 ms (t_{11}=-1.32, p=0.11) and 100 ms (t_{11}=-1.77 p=0.052,) while it remained non-significant for shorter SOAs. Furthermore, the effect of SOA was mainly observed on error trials (p= 0.006) and did not reach significance on correct trials (p=0.22). No such effect was observed for *unseen* trials.

The discussion of these results is made complicated by the many differences in the behavior of the Pe across experiments 1 and 2. At the very least, the results confirm the previous reported dissociation between the Pe and the ERN (Hewig, Coles, & Trippe, 2011; Nieuwenhuis, Ridderinkhof, Blom, Band, & Kok, 2001; Steinhauser & Yeung, 2010) by showing that the ERN can remain present even when the Pe vanishes (experiment 1) or, vice-versa, that the Pe can be present even in the absence of an ERN (*unseen* trials in experiment 2).

One possible, though admittedly highly speculative interpretation, is that the Pe reflects, at least in part, the degree of confidence in one's initial response, at the time at which it is emitted (perhaps due to the fact that we extract it to a short time window time-locked to the response). The fact that the Pe was not observed in Experiment 1, where time-pressure was stronger, RTs were overall faster, and responses were much less accurate, suggests that the level of evidence at the time of the response has an important impact on the subsequent amplitude of the Pe. Indeed, in experiment 2 where first-order and second-order performance were above chance both for *seen* and *unseen* trials, the Pe was observed in both cases. Such a result is in accordance with previous findings showing that the Pe reflects the evidence-

accumulation process leading to performance-evaluation judgment (Steinhauser & Yeung, 2012). In particular, the presence of the Pe for *unseen* trials in the second experiment, although surprising, may be explained by considering that second-order judgments were indeed above chance in this condition.

Overall, the Pe results partially fit with the hypothesis that the Pe is linked to error signaling and confidence judgment (Dhar, Wiersema, & Pourtois, 2011; Hughes & Yeung, 2011; O'Connell et al., 2007; Steinhauser & Yeung, 2010). Nevertheless, problems remain, both in understanding why the Pe was barely detectable on *seen* trials in experiment 1 (when errors were nearly always detected), and why the Pe was of equivalent size for *seen* and *unseen* trials in experiment 2 (while metaperformance differed widely). More research will be needed to understand these striking discrepancies with the confidence model.

References

Carbonnell, L., & Falkenstein, M. (2006). Does the error negativity reflect the degree of response conflict? *Brain Res*, *1095*(1), 124–130. doi:S0006-8993(06)01026-2 [pii] 10.1016/j.brainres.2006.04.004

Dhar, M., Wiersema, J. R., & Pourtois, G. (2011). Cascade of Neural Events Leading from Error Commission to Subsequent Awareness Revealed Using EEG Source Imaging. (A. Sirigu, Ed.)*PLoS ONE*, *6*(5), e19578. doi:10.1371/journal.pone.0019578

Hewig, J., Coles, M., & Trippe, R. (2011). Dissociation of Pe and ERN/Ne in the conscious recognition of an error. …, 1–7. doi:10.1111/j.1469-8986.2011.01209.x

Hughes, G., & Yeung, N. (2011). Dissociable correlates of response conflict and error awareness in error-related brain activity. *Neuropsychologia*, *49*(3), 405–15. doi:10.1016/j.neuropsychologia.2010.11.036

Nieuwenhuis, S., Ridderinkhof, K. R., Blom, J. H., Band, G. P. H., & Kok, A. (2001). Error-related brain potentials are differentially related to awareness of response errors: evidence from an antisaccade task. *Psychophysiology*, *38*(5), 752–760.

O'Connell, R. G., Dockree, P. M., Bellgrove, M. A., Kelly, S. P., Hester, R., Garavan, H., Robertson, I. H., et al. (2007). The role of cingulate cortex in the detection of errors with and without awareness: a high-density electrical mapping study. *Eur J Neurosci*, *25*(8), 2571–2579. doi:EJN5477 [pii] 10.1111/j.1460-9568.2007.05477.x

Steinhauser, M., & Yeung, N. (2010). Decision processes in human performance monitoring. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *30*(46), 15643–53. doi:10.1523/JNEUROSCI.1899-10.2010

Steinhauser, M., & Yeung, N. (2012). Error awareness as evidence accumulation: effects of speed-accuracy trade-off on error signaling. *Frontiers in human neuroscience*, *6*(August), 240. doi:10.3389/fnhum.2012.00240

CHAPTER 5

Article 2 : Decoding the dynamics of action, intention, and error-detection for conscious and subliminal stimuli

---

5.1 Introduction to the article

5.1.1 Context and goal of the study

We established in our previous study that the ERN was present only in conscious trials while some remaining metacognitive information related to confidence in the response could be extracted from non-conscious trials. What then might qualitatively differ between conscious and non-conscious trials that allow the ERN to be triggered in one condition and to be absent in the other?

It has been proposed that the ERN reflects either the conflict or the mismatch between two representations: the representation of the executed motor action and the representation of the correct response (Falkenstein et al., 2000; Yeung et al., 2004). Independently of the validity of both theories as well as the details of their models, both suppose that a representation of the correct response exists in brain activity even when making an error. Indeed, it is difficult to imagine a model allowing the detection of errors with near certainty that does not rely on a representation of the correct response in order to monitor the accuracy of decisions. However, this hypothesis makes a very strong prediction: for every detected error there should be a representation in brain activity indexing that it is the opposite motor response that is required. In other words, for each trial, information about the correct/required action should be present, independently of the ongoing action, even when it is erroneous. Such a view could explain the absence of ERN in non-conscious trials. According to this hypothesis, if such a representation failed to be established, it is impossible to determine with high accuracy the performance on a given trial. Therefore, if such representation could not be established in non-conscious trials, this could explain the absence of the ERN.

5.1.2 Experiment

The main question that we address in the present study is whether it is possible to isolate in brain activity a representation of the correct response, even when we are making an error. To investigate

this question we used multivariate pattern analysis on M/EEG data in order to obtain the dynamics of accumulation of evidence for different stages of stimulus processing and action monitoring. Using the same masking paradigm in which subjects assessed the visibility of the target stimulus on a trial-by-trial basis, we separated trials according to subjective visibility in order to address two main questions:

1. Is it possible to decode a representation of the correct response independently of the ongoing motor action in *seen* and in *unseen* trials?

2. Does this representation influence the subsequent error detection process? In particular does the level of evidence concerning the correct response and the moment it emerges in time correlate with how well and when we are able to detect our errors?

Putting aside the question of consciousness, to what brain patterns might representation of the correct/required response correspond? We know that at the time of the response, brain activity is dominated by signal linked to motor preparation and somato-sensory feedback. In the time following the response, error-related activity is also very strongly captured by neuroimaging techniques such as MEG and EEG. However, our prediction is that the computation of the correct/required response should be distinct from these two processes, corresponding to a third distinguishable pattern of activity (Figure 5.1). As this representation should be common to correct and error trials, it is important to unsure that both types of trials, errors and corrects, are used equally by the decoder (Figure 5.1).

Considering a decoding approach, how can we decode a common representation of the required response in correct and error trials? One approach could be to train a classifier on correct trials to discriminate the trials according to the required response and then try to generalize this classification to error trials. However, as can been seen on Figure 5.1, training a classifier this way would end up in it learning to classify trials according to the motor response. Therefore, the most probable outcome of generalization would be for the classifier to also classify error trials according to the motor response, rather than the required response. On the opposite, if the we train a decoder to classify trials according to the required response only on trials corresponding to one motor response (for example left response trials), we will end up learning to decode whether the response was in fact an error or on the contrary was correct (see Figure 5.1). Therefore a generalization approach cannot be valid in this case. The only way to teach a decoder to decode the required response independently of the executed motor response is to train it on both error and correct trials, giving as much importance to both types of trials in the fitting procedure.

Additionally, a crucial element in decoding the required response from error and correct trials is to use a linear decoder. To understand this point, let us consider that our data are in a three dimensional space (Figure 5.1) and that for each data point (represented as a dot or a star on the figure), three coordinates are available. One coordinate codes for the motor response (x-axis), one coordinate codes for the required response (z-axis) and one coordinate codes for the accuracy of the response (y-axis). We obtain four clouds of points corresponding to the four possible types of trials ($R_{left/left/correct}$,

		ACTUAL RESPONSE	
		LEFT	RIGHT
REQUIRED RESPONSE	LEFT	Correct	Error
	RIGHT	Error	Correct

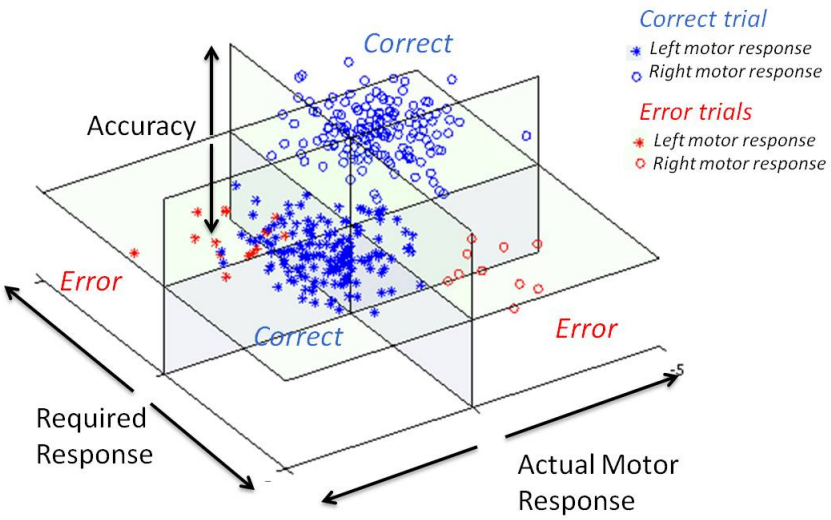


Figure 5.1: The two-by-two design of the experiment and the three decoded dimensions. The top table shows the four types of trials of our design, according to the actual motor response, the required response, and the accuracy. The bottom graphs shows how samples of simulated data are distributed along these three dimensions. The x-axis codes for the actual motor response, the y-axis code for the accuracy and the z-axis code for the required response. Blue dots correspond to correct trials and red dots to error trials. Filled dots correspond to right motor response and full dots correspond to left motor response.

$R_{left/right/error}$, $R_{right/left/correct}$ and $R_{right/right/correct}$). It is therefore possible to train three classifiers to decode respectively each dimension: a classifier for the motor response (x dimension), a classifier for the required response (z-dimension) and a classifier for the accuracy (y dimension).

Interestingly, as accuracy depends both on the actual and the required response, it can be deduced from the two others. Therefore, for non-linear classifiers, only two dimensions are necessary to classify trials according to the third dimension. In other words, if we consider again our three classifiers for actual response, required response and accuracy, it is sufficient to find two classifiers for the first and the third classification problem, to obtain by combining them in a non-linear way a classifier for the remaining dimension.

On the contrary, if we use a linear classifier, the three dimensions become necessary for our three classifiers. An intuitive way to see this is to put aside one dimension (for example the accuracy dimension represented along the y-axis) and project the data in a two dimensional space (in the case of accuracy, observe the data from above). When the data are in this two-dimensional space, no line can be

found that separates the data according to the third dimension. Indeed, the use of a linear classifier forces the use of an orthogonal dimension to the two others to classify trials according to the third dimension. In other words, decoding of the required response cannot rely on the decoding of the motor response or of the accuracy if linear classifiers are used. This demonstrates that if we find a linear classifier for the required response using both error and correct trials, it means a common pattern of activity corresponding to the representation of the correct/intended action exist in the brain, which is distinct from the motor response or the error-related activity.

5.1.3 Summary of the results

To summarize our results we found that:

- A representation of the intended action exists in brain activity as it is possible to train a classifier to predict the required response, independently of the actual motor response.
- Such a representation is present only in conscious trials.
- Error detection seems to result from the comparison between the required and the executed response, its certainty and time of occurrence being determined by the emergence of these two pieces of information.

We propose that our finding can be explained by a dual-route model for error detection in which accuracy is determined by comparing the outputs of two distinct routes: a fast non-conscious route that triggers motor action and a slow conscious route that computes intention.

5.2 Article

Charles, L., King, J.-R. & Dehaene, S. 2013 Decoding the dynamics of action, intention, and error-detection for conscious and subliminal stimuli. In revision

Decoding the dynamics of action, intention, and error-detection for conscious and subliminal stimuli

Lucie Charles [a, b, c] , Jean-Rémi King [a, b, c] and Stanislas Dehaene [a, b, c, d]

a INSERM, U992, Cognitive Neuroimaging Unit,
CEA/SAC/DSV/DRM/NeuroSpin
Bât 145, Point Courrier 156
F-91191 Gif/Yvette, France

b CEA, DSV/I2BM, NeuroSpin Center
Bât 145, Point Courrier 156
F-91191 Gif/Yvette, France

c Univ Paris-Sud, Cognitive Neuroimaging Unit,
Bât. 300 - 91405 Orsay cedex

d Collège de France,
11, place Marcelin Berthelot
75231 Paris Cedex 05, France

Corresponding author: Lucie Charles

INSERM-CEA Cognitive Neuroimaging unit
CEA/SAC/DSV/DRM/NeuroSpin
Bât 145, Point Courrier 156
F-91191 Gif/Yvette, FRANCE
Tel: +33 1 69 08 99 74
Fax: +33 1 69 08 79 73
lucie.charles.ens@gmail.com

Number of pages: 42

Number of figures: 6

Number of words for Abstract: 250

Number of words for Introduction: 491

Number of words for Discussion: 1335

Acknowlegments

This project was supported by PhD grants from the Direction Générale de l'Armement (DGA, Didier Bazalgette) and the Fondation pour la Recherche Médicale (FRM) as well as a senior grant of the European Research Council to S.D. (NeuroConsc program). The NeuroSpin MEG facility was sponsored by grants from INSERM, CEA, the Fondation pour la Recherche Médicale, the Bettencourt-Schueller foundation, and the Région île-de-France. F.V.O. is a Postdoctoral Fellow of the Research Foundation– Flanders (FWO-Vlaanderen). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors declare no competing financial interests.

We are grateful to the NeuroSpin infrastructure groups, in particular to the doctors Ghislaine Dehaene-Lambertz, Andreas Kleinschmidt, Caroline Huron, Lucie Hertz-Pannier and the nurses Véronique Joly-Testault and Laurence Laurier, for their support in subject recruitment and testing; Virginie van Wassenhove, Marco Buiatti, Leila Rogeau, Etienne Labyt and the NeuroSpin MEG team for their technical help; and Bertrand Thirion, Gaël Varoquaux, Alexandre Gramfort and Fabian Pedregosa for their assistance with decoding methods.

Abstract

How do we detect our own errors, even before we receive any external feedback? One model hypothesizes that error detection results from the confrontation of two signals: a fast and unconscious motor code, based on a direct sensory-motor pathway, and a slower conscious intention code that computes the required response given the stimulus and task instructions. To test this theory and assess how the chain of cognitive processes leading to error detection is modulated by consciousness, we applied multivariate decoding methods to single-trial magneto- (MEG) and electro-encephalographic (EEG) data. Participants performed a fast bimanual number comparison task on masked digits presented at threshold, such that about half of them remained unseen. By using both erroneous and correct trials, we designed orthogonal decoders for the actual response (left or right), the required response (left or right), and the response accuracy (correct or incorrect). While perceptual stimulus information and actual response hand could be decoded on both conscious and non-conscious trials, the required response could only be decoded on conscious trials. Moreover, whether the current response was correct or incorrect could be decoded only when the target digits were conscious, at a time and with a certainty that varied with the amount of evidence in favor of the correct response. These results are in accordance with the proposed dual-route model of conscious versus non-conscious evidence accumulation, and suggest that explicit error detection is only possible when the brain computes a conscious representation of the desired response, distinct from the ongoing motor program.

I. Introduction

Performance monitoring is a key function of the cognitive control system. When speed is emphasized over accuracy, we often commit a large numbers of errors that we are nonetheless able to correct and detect in a fast automatic manner (Rabbitt, 1966; Gehring et al., 1993). But how can the very same system that commits an error detect it? According to some models of cognitive control (Norman and Shallice, 1986), decision and motor control are organized as a hierarchy in which higher-level conscious and intentional processes attempt to monitor performance (Posner and Rothbart, 1998) but sometimes arrives too late to modulate ongoing actions (Norman, 1981; Rabbitt, 2002).

In particular, the dual-route model for conscious and non-conscious decision making (Del Cul et al., 2009) hypothesizes that whenever we have to produce a motor response to some stimulus, two parallel routes (Figure 1B) simultaneously accumulate evidence from the sensory input: a fast non-conscious sensory-motor route, and a slower but more accurate conscious route. Crucially, when instructions emphasize speed over accuracy, responses may frequently be emitted by the unconscious route, before the slower conscious route emits its more conservative judgment. Any discrepancy between the outputs of these two routes indicates that an error was committed, signaling a "mismatch" (Coles et al., 2001) or a conflict (Yeung et al., 2004) between actual and intended actions.

Here, we aimed at testing several prediction of this dual-route model. First, this model predicts that at the same time as the subject is making an erroneous action, for instance clicking on the left-hand button, his or her brain should contain a distinct representation of the required correct action (i.e. clicking right). Second, this signal should only be present on conscious trials, coding for the conscious intention of the subject. Third, we should be able to predict the brain's capacity for spontaneous error detection by the strength and timing of the discrepancy between the action and intention codes.

In the present study, we investigated this question by attempting to decode, from single-trial brain activity, the chain of cognitive processes linked to action monitoring and determine how each stage was modulated by consciousness access. Decoding techniques, such as Multivariate Pattern Analysis (MVPA) have proven powerful to isolate precise cognitive processes in brain activity (Norman et al., 2006) and distinguish the sequence of processing of a specific task (Bode and Haynes, 2009). Here, we applied these techniques to high temporal-resolution magneto- (MEG) and encephalographic (EEG) recordings while participants performed a speeded number-comparison task on a masked digit, and reported on each trial their subjective perception (*seen* / *unseen*) of the target (Charles et al., 2013). By training decoders to classify trials according to four different features (stimulus position, actual motor action, required motor action and accuracy), separately on *seen* and *unseen* trials, we assessed how subjective visibility modulated perception, action, intention and error detection. We then tested the prediction of the dual-route model that error-detection result from the comparison of actions and intentions.

;

II. Methods

A. Participants.

13 volunteers (with normal or corrected-to-normal vision) were tested in this MEG/EEG experiment. Event Related Potential and Event Related Fields of these data have been partially reported elsewhere (Charles et al., 2013). As the present within-subject decoding analysis necessitated a large number of error trials with both left and right motor responses, participants were excluded from the analysis if they did not have at least 20 error responses on both type of motor responses. Six participants (3 men, 3 women) had sufficient numbers of trials in all of the conditions and were kept for analysis. This small number was compensated by the fact that we systematically examined the within-subject significance of decoding scores, thus obtaining, for each question we raised, six within-subject replications as well a between-subject non-parametric test.

B. Design & Procedure

The paradigm of this experiment is described in detail in Charles et al (Del Cul et al., 2007; Charles et al., 2013). Briefly, a target-stimulus (the digit 1, 4, 6, or 9) appeared on a white screen for 16 ms at one of two positions (top or bottom, 2.29 degrees from fixation), with a pseudo-random 50% probability. After a variable delay, a mask appeared at the target location for 250ms. The mask was composed of four letters (two E and two M, see Figure 1A) tightly surrounding the target stimulus without superimposing or touching it. The stimulus-onset asynchrony (SOA) between the onset of the target and the onset of the mask was varied across trials. Five SOAs were randomly intermixed: 16, 33, 50, 66 and 100ms. In one sixth of the

trials, the target number was replaced by a blank screen with the same duration of 16ms (mask-only condition), allowing us to study visibility ratings when no target was presented.

Participants primarily performed a speeded forced-choice task of comparing the target number to the number 5. Responses were collected within 1000 ms after target onset with two buttons using the index of each hand (left button press = smaller-than-5; right button-press = larger-than-5 response). To induce errors, participants were instructed to respond as fast as they could just after the appearance of the target. Time pressure was increased by presenting an unpleasant sound (mean pitch: 136.2 Hz, 215 ms duration) 1000 ms after target presentation whenever response time exceeded 550ms.

At the end of each trial, after another delay of 500 ms, participants were requested to provide two subjective answers with no time pressure. First they had to indicate whether they saw the target number or not (visibility task). Second, they had to report whether they thought they had made an error or not in the number comparison task (performance evaluation task). For both subjective responses, words corresponding to the two responses (*seen*/*unseen* and *error/correct*) were displayed on the screen and subjects had to use the corresponding-side buttons to answer. The words were presented at randomized left and right locations (2.3 degrees from fixation) to ensure that subjects did not use an automatized button-press strategy.

The experiment was divided in blocks of 96 trials, with 16 trials by SOA condition in which each digit was presented at the two possible target locations (Top/Bottom). Each participant performed six or seven blocks during M/EEG recording. For experiment 1, in order to achieve fast responses, participants were given a training session before the actual recording. They first received five minutes of training during which the target stimulus was not masked. Participants then performed three pre-recording blocks of the actual experiment in order to check that overall performance was suitable for MEG/EEG recording.

C. Simultaneous EEG and MEG recordings.

Simultaneous recording of MEG and EEG data was performed. The MEG system (the Elekta-Neuromag) comprised 306 sensors: 102 Magnetometers and 204 orthogonal planar gradiometers (pairs of sensors measuring the longitudinal and latitudinal derivatives of the magnetic field). The EEG system consisted of a cap of 60 electrodes with reference on the nose and ground on the clavicle bone. Six additional electrodes were used to record electrocardiographic (ECG) and electro-oculographic (vertical and horizontal EOG) signals.

A 3-dimensional Fastrak digitizer (Polhemus, USA) was used to digitize the position of three fiducial head landmarks (Nasion and Pre-auricular points) and four coils used as indicators of head position in the MEG helmet, for further alignment with MRI data. Sampling rate was set at 1000 Hz with a hardware band-pass filter from 0.1 to 330 Hz.

D. MEG/EEG Data Preprocessing.

MEG data were first processed with MaxFilter^{TM} software using the Signal Space Separation algorithm. Bad MEG channels were detected both automatically and manually, and were subsequently interpolated. Head position information recorded at the beginning of each block was used to realign head position across runs and transform the signal to a standard head position framework.

To remove the remaining noise, Principal Component Analysis (PCA) was applied to regress out the stereotypical physiological artifacts. First, artifacted time periods were detected on the electro-occulogram (EOG) and electro-cardiogram. Second, data were averaged on the onset of each blink and each heart beat separately and PCA was performed separately for each type of sensor. Then, one to three of the first components characterizing the artifact were manually selected to be further removed.

Data were then entered into Matlab software and processed with Fieldtrip software (http://fieldtrip.fcdonders.nl/). An automatic rejection of trials based on signal discontinuities (all signal above 30 and 25 standard deviations in 110-140 Hz frequency range) was performed. A low-pass filter at 30 Hz was then applied as well as a baseline correction from 300 ms to 200 ms before target stimulus onset.

An additional process was applied to the data used to decode stimulus position. Since the mask stimulus was presented at the same position as the target digit, we subtracted out the activity evoked by the mask, in order to minimize the information provided by the mask location and decode only the information about the masked target. To do so, we first aligned each trial on the mask onset. We averaged separately the trials for which no target was presented, corresponding to the mask alone condition. We then subtracted from the rest of the data this mask-related activity and realigned those subtracted data on target onset (Del Cul et al., 2007). As stimulus-position was not relevant for the other decoded categories, we did not apply this method to other decoding stages.

E. Decoding Analysis

The support vector machine (SVM) method was used to decode different stages of perceptual decision, from stimulus encoding to performance detection. Briefly, linear classifiers such as SVM allow to discriminate, on a single-trial basis, two conditions based on their pattern of activity across trials (Chang and Lin, 2011). This is achieved by finding a hyperplane separating the two classes of trials along the dimensions given to the decoder (such as sensors or time). We tested eight different decoders with binary SVM classification for testing several perceptual stages. Crucially, we split the initial dataset according to visibility reports and tested on each subset of trials how well the classifier could discriminate each of the four following conditions: stimulus position, *top* versus *bottom*; actual motor

response, *left* versus *right*; correct motor response, *left* versus *right*; accuracy, *error* versus *correct*.

Importantly, as several of these categories are based on responses of the participants, some of them had unbalanced number of trials. For instance, more responses were made with the right hand than with the left hand, probably because subjects who were in majority right handed were more prompt to answer with their dominant hand. Furthermore, difference between categories could be also partially confounded with different numbers of trials in each subset. For instance, more errors were made when making a right-hand response, for similar reasons, therefore making more erroneous fast guesses with this response hand. These confounds present a problem for the decoding approach: for instance, when trying to decode error versus correct trials, we might obtain better than chance results simply because we are decoding left versus right responses.

To counteract these biases, we applied sample-weights in order to equalize the weight of the trials entering in the classification and belonging to each cell of the potentially confounding category. As noted earlier, this required selecting participants who had more than 20 trials in each trial subcategory, therefore securing that each subcategory was sufficiently populated. Sample weights were applied to each trial according to the number of trials in the sub-category to be controlled for. Each sample weight was computed using the following formula:

$$w_{categ} = n_{tot}/(4*n_{categ})$$

where w_{categ} designate the weight of all of the samples in the subject cagerory; n_{tot} designate the total number of trials provided to the classifier and n_{categ} designate the total number of trials in this subcategory. Note that total sum of weights across all trials was equal to n_{tot}, similarly to the case where a weighting of 1 is applied to each trial. The subcategories of trials that we controlled for were the required motor response (left or right) for the position decoder

and the actual motor response decoder, and the actual motor response (left or right) for the correct response and the accuracy decoders respectively. Sample weights were entered as parameters in a linear-kernel SVM implemented by SciKit-Learn toolbox (Pedregosa et al., 2011). For the decoding of the required response for which the results were the most crucial, we performed an additional analysis step to unsure that the decoder was not biased by the unbalanced number of trials among classes. In particular, as decoding the required response is identical to decoding the actual response in correct trials and these trials were more numerous than error trials, the decoder could simply end up separating the trials according to the motor action (left versus right). To unsure this was not the case, we separated the data according to the response hand and verified that within each subset, the decoder performed above chance in classifying the trials according to the required motor response.

For each participant, MEG/EEG pre-processed data were entered into the classification pipeline. Importantly, we used two types of decoders, differing in the features that the decoder was trained on. In the first case, both time and space were used as decoding features and the decoder was provided with the entire trial time-window (0-800 ms after stimulus presentation). In this case, the decoder learned to decode in a high dimensional space with a total of $n_{time\text{-}point}$*$n_{channel}$ dimensions. In the second case, we trained a different decoder for each time point, using as a feature only the spatial dimension ($n_{channel}$ dimensions). For the first type of decoder, we obtained only one classification measure for each trial. In the second case, we could reconstruct for each trial the entire time course of classification accuracy, allowing us to study more precisely the dynamics of the related cognitive process.

All decoding stages (including normalization of the MEG/EEG data) were fitted within the cross-validation loop on the training sets only. To obtain a valid training/testing datasets, we used stratified k-folding method, according to the number of trials in each subcategory (as described above). The data were split in 7 folds, each fold being composed of

a testing set of 1/7 of the trials and a training set of 6/7 of the trials, with the same proportion of trials coming from each subcategory.

Each training dataset underwent a first stage of feature selection, using an ANOVA to keep only the 50% most informative features. The remaining feature*trial training data were then rescaled using the mean of z-score transformation. This method allows for all the EEG and MEG channels, which are recorded in different physical units, to be put on the same scale and used properly by the classifier. The penalization parameter of the algorithm was then estimated by mean of a grid search by nested cross-validation applied within each training dataset (2 stratified k-fold) and the best hyperplane was retrieved. Finally, we fitted a cumulative probability distribution function on the decision function of the training dataset using Platt's method (Platt, 1999), allowing us to obtain for each trial, not just a discrete output label, but a continuous value bounded between 0 and 1, representing the classifier's estimate of the probability to belong to the first class. Then, exactly the same feature selection and scaling parameters obtained from training dataset were applied to the testing dataset and the obtained classifier was applied to the test trials, allowing us to obtain a cross-validated classification measure for each of the test trials. We ensured that we applied all multivariate classification guidelines outline in Lemm et al. (2011) in order to minimize classification bias and avoid circular analyses that could result in overfitting the data.

F. Statistical analysis

1. Within-subject classification scores

Classification scores across trials were estimated for each subject with a receiver-operative curve (ROC) analysis applied to the obtained classification probabilities, and were summarized by the area under the curve (AUC) values. The ROC curve presents the true positive rate (the proportion of trials belonging to class A and classified as A, i.e. hits) as a

function of false positive rate (the proportion of trials belonging to class B and classified as A, i.e. false alarms) providing a measure of both sensitivity and specificity of the decoder. A diagonal ROC curve, which coincide with an AUC of 50%, corresponds to a situation where the number of hits and false alarms are equal, showing a chance level classification score. On the contrary, an AUC of 100% which corresponds to a ROC curve on the left upper bound of the diagonal, indicates a perfect positive prediction with no false positives and a perfect decoding score. Importantly, and unlike average accuracy AUC analysis provides an unbiased measure of decoding accuracy, robust to imbalanced problems and independent of the statistical distribution of the classes.

The classification AUCs were estimated for each subject for the decoders on the entire-trial duration (Figure 2, right column) and above-chance significance within- and across-subjects was computed by means of a non-parametric Wilcoxon rank sum test. Separatly, the AUC was computed for the obtained decoding time-series, separately for each time-point and was averaged across subjects. The middle columns in Figure 2 show the AUC time-series averaged across subjects for each decoder.

2. Within-subject time cluster analysis on decoder time-series

To determine the moments at which the decoders performed above chance, we computed within-subject statistics on the obtained trial time series. Using individual data, for each decoder, we used a cluster-based non-parametric with Monte Carlo randomization (adapted from Maris and Oostenveld, 2007) on the trial-by-trial time-series of decoding probabilities. This method allowed us to identify clusters of time-points in which time-series of the two learned classes present a significant difference while correcting for multiple comparisons. For each time-sample, p-values of the difference between the two decoded classes were first computed by means of a non-parametric Mann–Whitney U test. Clusters

were then identified by taking all dyads of time-samples adjacent in time with $p<0.05$. The final significance of the cluster was determined by computing the sum of AUC-values of the entire cluster, and comparing with the results of Monte-Carlo permutations (2000 permutations). Clusters were considered significant at corrected $p<0.05$ if the probability computed with the Monte-Carlo method was inferior to 5% (one-tailed test). The number of subjects presenting a significant cluster at each time-point is showed in Figure 2 at the bottom of each graph.

3. Regression analysis on single-trial amplitude

The dual-route model of error detection predicts that on each trial, evidence on the required response and the actual response are compared in order to determine the accuracy of the action. In other word, for a given trial, the amount of evidence that an error was made depends on the discrepancy between action and intention.

To evaluate this prediction, we used the actual and the required response entire-trial decoders as indices of the amount of internal information available on each trial about the action and the intention. As chance-level was not identical across subject, we normalized for each subject the trial-by-trial classification probability. We then transformed the obtained signal so that it would be centered on 0 and fluctuate between 1 and -1 (instead of 0 and 1, Figure 4A). This can be achieved by subtracting on a trial-by-trial basis the decoded probability of belonging to one of the two classes from the probability of belonging to the opposite class (Figure 4A): as the sum of the probability is equal to one, when the probability of belonging to one class is close to one, the subtraction will be close to either +1 or -1, while it will be close from 0 when the probability are at chance. We then computed the product of the two obtained indices, this measure giving us a trial-by-trial index of the discrepancy between action and intention. We then retrieved, for each trial, the output of the accuracy

decoder corresponding to the decoded trial-by-trial probability an erroneous motor response and correlated it to our index of congruity between action and intention.

Using robust linear regression (Holland and Welsch, 1977), we correlated for each subject the trial-by-trial indices obtained by multiplying the motor and intention decoder with the accuracy probability. A non-parametric test was then performed on the slope of the regression obtained across subjects. As negative values of the computed product should signal erroneous responses, we expected a negative correlation between the two measures, the smallest negative value being associated with the highest probability of decoding an error.

4. Regression analysis on single-trial timing

Another prediction of the dual-route model is that one should be able to determine the accuracy of its own response only when information is available on both the required response and the response actually made. This prediction implies that the latest obtained information either on the actual response or the correct response, should determine the moment at which an internal estimate of response accuracy can be emitted.

To test this prediction, we searched for a correlation between the time at which the accuracy decoder crossed a threshold and the moment when the latest of the action and intention decoders crossed their threshold. For each subject separately, we normalized the classification probability time-series according to the baseline (-100 to 0 before stimulus presentation) in order to obtain values centered on 0 and ranging from -1 to 1. To increase the signal-to-noise ratio, we converted the SVM probability time-series into a cumulative-sum time series (Figure 6A) and we extracted for each decoder the moment at which each time-series reached 50% of its mean final value across trials (Figure 6B). Trials which did not reach the threshold for any of the three decoders were excluded from the analysis, resulting on the selection of about half of the trials for which intention, action and accuracy could be

decoded with high performance. We then took the maximal value of the decision times for the actual response and the required response decoders and correlated it with the crossing of the threshold of the accuracy decoder. We then performed a non-parametric Wilcoxon signed-rank test on the betas across subjects.

III.Results

A. Decoding stages of stimulus processing

In order to isolate the effect of consciousness on the processing chain leading from the stimulus to the response and its evaluation, we separated different stages in a decision hierarchy, and we tested whether and when an experimental variable attached to each processing stage could be decoded from the single-trial brain activity, separately for conscious and non-conscious trials. Figure 2 depicts, for each decoder, the individual classification score (AUC, see Methods) over the entire trial window and the time course of the classification score, averaged across subjects.

1. Decoding early visual processes: stimulus position classifier

Figure 2C shows the result of the classification of stimuli position over the entire trial duration, for both *seen* and *unseen* trials. This analysis revealed that stimulus position could be decoded for each individual subject on both *seen* and *unseen* trials, with high accuracy. Non-parametric statistics showed that the decoder performed significantly above chance for each subject (Wilcoxon rank-sum test on classification probabilities, all $p < 10^{-4}$). The AUC was significantly higher than chance across subjects for both types of trials (Wilcoxon rank-sum test AUC > 0.5, n=6, both $p < 0.05$).

Considering the results of the decoding on each time point allowed us to determine precisely the dynamics of perceptual processing of the stimulus, in *seen* and in *unseen* conditions (Figure 2A-B). The peak of performance of the decoder was observed around 175 ms after stimulus presentation for *seen* trials and 130 ms for *unseen* trials. Within-subject statistical analysis revealed that both for *seen* and *unseen* trials, subjects presented a significant cluster starting around 75 ms after onset of the stimulus, and which lasted for at

least 450 ms. Interestingly, for 5 of the 6 subjects the time-window of significance lasted longer for conscious than for non-conscious trials.

We then performed a non-parametric test to determine if the overall difference between the two decoded classes was greater for *seen* compared to *unseen* trials. Non-parametric test across subjects revealed no statistical significance between the two ($p = 0.47$), suggesting that the performance in decoding stimulus position over the entire trial duration were not different in *seen* as compared to *unseen* trials.

These results suggest that it is possible to classify the stimulus position with as high accuracy both for *seen* compared to *unseen* trials, showing that early visual processing of the stimulus is largely unimpaired in non-conscious conditions.

2. Decoding motor decision: actual response decoder

We then turned to the motor response decoder. The aim of this analysis was to determine if it was possible for a decoder to learn which motor decision was made by the subject. According to our design, a left-hand action implies that the subject choose to respond that the stimulus was smaller than 5 while a right-hand action corresponded to a larger-than-5 response.

Figure 2D-F shows that the decoder performed significantly above-chance to determine if a left or a right motor response was produced on each trial, both for *seen* and for *unseen* trials. Again, analysis of the AUCs obtained from the decoding of the motor response over the entire time-window revealed that for each subject we were able to decoded the motor response better than chance both for *seen* and for *unseen* trials (Wilcoxon rank-sum test, all $p < 10^{-4}$, Figure 2F). Similarly, the AUC was significantly higher than chance across subjects both in conscious and non-conscious conditions (Wilcoxon rank-sum test, n=6, both $p < 0.05$), confirming that the motor response could be decoded with high accuracy in both

cases. Interestingly, comparison between *seen* and *unseen* trials revealed no statistical difference between the two, suggesting comparable decoding accuracy of the motor response in conscious and non-conscious conditions.

The time course of the decoding of the motor response (Figure 2D-E) revealed that decoding accuracy increased linearly from approximately 120 ms after stimulus presentation both for *seen* and *unseen* trials. For five out of six subjects, the earliest significant difference between left and right responses was observed at 240 ms after stimulus presentation. Decoding accuracy reached a plateau around the average time of the actual key press (365 ms and 366 ms respectively for *seen* and *unseen* trials). The maximal peak was observed around 425 ms for *seen* trials and 365 ms for *unseen* trials, slightly later than the mean RT across subjects. In summary, this analysis revealed that actual motor response could be decoded with very high accuracy both in conscious and in non-conscious conditions.

3. Required Response decoder

One of the main goals of this study was to test whether it is possible to decode, from the time course of brain activity, the presence of a higher-order representation of the required response. We predicted that, on top of the representation of the actual ongoing motor program, there might be a distinct representation of the intended response. On the majority of trials where the response is correct, the intended and actual responses coincide. However, whenever subjects commit an error, the dual-route model predicts that their brain contains a distinct representation of the response that would have been correct. Thus, this neural code should encode the response that should have been made by the subjects, independently of the response that they actually make.

To test this idea, we trained a decoder to classify trials according to the required response, regardless of the actual motor response on the same trial. Importantly, in order to teach the decoder the proper class, we weighted equally the erroneous and correct trials. Since errors were overall much less frequent than correct trials, we used a weighting technique that ensured that both errors and correct trials were equally used in training the decoder (see Methods), thus removing the correlation between intended and actual responses.

On *seen* trials, decoding over the entire time-window revealed that we were able to decode the required response for each subject (Wilcoxon rank-sum test, all $p < 0.005$, Figure 2I). Analysis across subjects revealed that the average AUC was significantly above chance (Wilcoxon rank-sum testn n=6, $p< 0.05$). However, for *unseen* trials, we were not able to decode the required response. Analysis of the decoding results showed that the classifier performed at chance for all subjects (Wilcoxon rank-sum test, all $p > 0.35$) except for one subject for which the classifier performed significantly below chance (Wilcoxon rank-sum test, $p < 10^{-4}$, Figure 2I). Similarly, average decoding score across subjects did not differ from chance (Wilcoxon rank-sum test, n=6, p=0.35). This resulted in a significant effect of visibility on the decoding scores across subject (Wilcoxon rank-sum test, n=6, $p < 10^{-3}$).

When training the decoder on each time-sample (Figure 2G-H), within-subject statistical analysis revealed a significant cluster for all subjects in the *seen* condition (Figure 2G). Three subjects presented an identical significant temporal cluster between 350 ms and 750 ms after stimulus presentation, while the remaining subjects presented shorter period of significance in this time-window. Interestingly, decoding performance varied in time across subjects, some subjects presenting above-chance decoding accuracy only starting on average at 500 ms after stimulus presentation. No such decoding was possible for *unseen* trials (Figure 2H). Cluster-level significance was not achieved for most of the subjects. For one subject, a 10ms time-window of significance was found, unlikely to reflect a solid effect (subject 6,

700-710 ms after stimulus onset). For another subject, a more sustained cluster was found but at a time that is unlikely to be meaningful (subject 5: 925-960 ms after stimulus onset). Therefore, these results suggest that a representation of the required response can be decoded from brain activity in *seen* trials, but that not enough information is available on *unseen* trials for the classifier to extract this representation.

As the decoding of the required response was performed on highly unbalanced datasets, where correct trial were more numerous than error trials, we verified that the decoder on *seen* trials was not simply picking up the motor activity on correct trials. Thus, we separated the trials according to the actual motor response (left versus right). Crucially, we verified that, within each such subset, the decoder performed above chance in classifying the trials according to the required motor response (Figure 3). When considering the entire time-window (Figure 3B), the decoder performed above-chance for 4 subjects considering the left motor responses and for 5 considering the right motor response (see Table 1). The average intention decoding AUCs across subjects were significantly above chance both for right (Wilcoxon rank-sum test AUC > 0.5, n=6, $p = 0.016$) and for left motor responses (Wilcoxon rank-sum test AUC > 0.5, n = 6, $p = 0.03$).

Furthermore, analysis of the decoding scores showed a simultaneous peak of decoding accuracy around 580 ms for both motor responses (Figure 3A). This analysis confirms that the intention decoder learned to classify trials according to the required response, independently of the actual motor response made by the subject.

4. Accuracy decoder

We then determined whether our recordings contained decodable single-trial information about the accuracy of the motor decision, separately for *seen* and *unseen* trials. The dual-route model postulates that in order to determine the accuracy of their decisions,

participants compare their actual motor response to the response that they should have made, and evaluate the discrepancy between these two internal representations. As we were not able to decode the representation of the required response on *unseen* trials, the model predicted that we should also not be able to decode accuracy on these trials. That is indeed what we found. Considering the entire time-window, we were able to decode with high performance the accuracy of the response at a trial-by-trial level for all 6 subjects on *seen* trials (Wilcoxon rank-sum test, all $p < 10^{-4}$), resulting in an above-chance classification score across subjects (Wilcoxon rank-sum test, n=6, $p<0.05$). Importantly, we were not able to decode the accuracy of the motor response on *unseen* trials except for one subject (Wilcoxon rank-sum test, all $p = 0.03$) resulting in a chance-level decoding score across subjects (Wilcoxon rank-sum test, $n = 6$, all $p = 0.08$). This resulted in a significant effect of visibility on the decoding scores across subjects (Wilcoxon rank-sum test, n=6, $p < 10^{-3}$).

Considering the decoding analysis for each time-sample, the peak of decoding performance on *seen* trials was reached at a latest time-window than for previous action and intention decoders, around 600 ms after stimulus presentation. On *unseen* trials, decoding scores remained at chance over the entire time-window.

Following the dual-route prediction model, our result therefore suggest that the brain encodes a representation of response accuracy that can be decoded with high accuracy on conscious trials, but that on non-conscious trials, when no information is available on the required response, the accuracy of the motor response cannot be predicted.

5. Analysis of the effect of visibility on early versus late processing stages.

Our result suggest that only the early stages of processing of the stimulus, containing either visual or motor activity, can be decoded equally well on conscious and non-conscious

trials, while higher-order representations of the goal of the action and its accuracy are available only in conscious conditions. To support this, we performed an ANOVA separately with visibility and decoder type (perceptual & motor versus intention & accuracy) as main factors. A significant interaction ($F_{1,47}$ = 21.07; p = 10^{-4}) revealed that, indeed, while early stages could be decoded with equal performance in conscious and non-conscious conditions, late stages could be decoded only in conscious trials.

B. Trial-by-trial test of predictions of the dual-route model

1. Congruity between action and intention correlates with the strength of error detection

The dual-route model states that if no representation of the required response is available, as seems to be the case in the *unseen* condition, then the accuracy of one's performance cannot be determined. A related prediction is that trial-by-trial variation in the amount of evidence, concerning either the required response or the actual response, should be predictive of the amount of evidence concerning decision accuracy. In particular, the more evidence one has on what the required response is, the better one can determine whether one's performance is correct or not.

To test this prediction, we focused on seen trials for which it was indeed possible to decode both intended and actual motor responses. We collected the trial-by-trial classification probabilities computed by the three main decoders and used them as indices of the amount of evidence available for the required response, the actual motor response and the accuracy (see Methods). Our main goal was to determine whether for each trial, the discrepancy between action and intention predicted the decoding of accuracy.

We computed for each trial an intention index and a motor index varying from -1 to 1 across trials, and coding for the amount of information that this trial contained, respectively, about the intended response and the motor response (Figure 4B; -1 corresponds to a sure Left response and +1 to a sure Right response). According to the dual-route model, the product of the intention and action indices, which evaluates their congruency, should predict the accuracy of the response. If the product is positive, it means that the action and the intention are congruent, and the actual motor response is therefore likely to be correct. On the contrary, if the product is negative it means that intention and action vote in favor of different responses, and the actual motor response is likely to be incorrect. Note that if one of the indices is close to 0, i.e. no information is available either on the action or on the intention, the product is also close to 0, so the model predicts that the accuracy of the response cannot be predicted. Across trials, accuracy evidence should therefore be correlated with the product of action and intention indices.

After transforming the classification probability of actual motor response and required response into signed indices of action and intention strength (Figure 4A), we computed for each trial the product of these two indices, obtaining a measure, for each trial, of the congruity between intention and action (see Methods). We then retrieved from the accuracy decoder the estimated probability that the response was erroneous on the same trial. Finally, we tested whether these two measures were correlated, as predicted by the dual-route model. As the obtained indices were signed values, we expected a negative correlation between the two measures: a negative product indicated a discrepancy between action and intention, and therefore a greater probability of error (Figure 4B).

Figure 4C depicts this correlation for each subject. Linear regression was performed for each subject and we tested whether the slope differed from 0. All regression slopes were negative (see Figure 4C) and Wilcoxon rank sum test on the slopes across subjects confirmed

that the average slope was significantly different from 0 ($n = 6$, $p = 0.016$). These findings indicated that indeed, trial-by-trial fluctuations in the congruity between intention and action signals correlated with fluctuations in the strength of error representation in the participants' brain, as predicted by the dual-route model.

2. The timing of error detection correlates with the slowest of the internal codes for action and intention

Another prediction of the dual-route model is that the timing of error detection should be predictable from the timing of the computation of the action and intention codes. To investigate more precisely this question, we realigned the obtained decoding time-series on the onset of the response, in order to gain a clearer view of how the dynamics of error detection varied with the timing of the response. Figure 5 depicts the time-courses of the classification scores realigned on the onset of the motor response in *seen* trials. Above-chance decoding of the motor response (Figure 5A) and the required response (Figure 5B) occurred prior to the onset of the actual key press. Classification performance was significantly better than chance in the time-window of -150 to -100 ms before the key press for the actual response decoder (Wilcoxon rank-sum test, $n = 6$, $p < 0.05$) and -100 to -50 ms for the required response decoder (Wilcoxon rank-sum test, $n = 6$, $p < 0.05$). Crucially, decoding of the accuracy was possible immediately after this point, in the time-window just preceding the motor response (-50 to 0 ms before key-press, Wilcoxon rank-sum test, $n = 6$, $p < 0.05$, Figure 5C), suggesting that error detection followed the computation of the actual response.

These results suggest that error detection immediately follows the erroneous motor action. However, when speed pressure is imposed, the dual-route model predicts that a motor response may be emitted early on, before a clear intention has been computed from the stimulus. In this case, error detection should only be possible once the intention is determined.

Overall, the timing of error detection should vary on a trial-by-trial basis according to the availability of both intention and action signals, whichever comes last.

To test this prediction, we computed for each trial the moment at which each of the three decoders (action, intention and accuracy) crossed a given threshold (see Methods). We therefore obtained for each trial three time measure T_{int}, T_{act} and T_{acc} corresponding respectively to the timing of intention, action and accuracy detection (Figure 6A). We then tested how these times correlated with one another.

We first verified whether our action time index, T_{act}, correlated with the actual trial-by-trial reaction time (RT). This was indeed the case: the slope of a linear regression was significantly greater than 0 across subjects (Wilcoxon Rank Sum test, $p = 0.016$).

As only the latest event between action and intention should determine when one can detect making an error, we then computed, for each trial, the maximum value between T_{int} and T_{act} and correlated it with T_{acc} as shown in Figure 6A. None of the regression reached significance at the single-subject level (all $p>0.05$). However, a non-parametric test on the slope of the regression across subjects revealed a significant positive correlation (Wicoxon rank sum test, $p = 0.016$, Figure 6B), suggesting a correlation between the timing of performance detection and the timing of action and intention, as predicted by the dual-route model.

IV. Discussion

We showed that M/EEG signals contain decodable information on the correct motor response, independently of the ongoing motor plan. Such information was present only on *seen* trials and not on *unseen* trials, while lower-level perceptual information and motor action were decodable on both types of trials. These findings suggest that, when the stimulus is masked below the threshold for conscious access, the brain is unable to compute a clear representation of the required action for that stimulus given the task instructions. Furthermore, the accuracy of the motor decision was also decodable from conscious trials only, with a magnitude and at a point in time correlating with the information decodable about the actual and the required action. These results fit with the prediction of the dual-route model of error detection, according to which accuracy can be determined, on conscious trials only, by comparing the output of two distinct cortical routes for conscious and non-conscious processes, which compute respectively intention and action.

The crucial finding of this study is that for conscious trials, a representation of the required response can be decoded in brain activity, independently of the ongoing motor action. This finding builds upon our previous work (Charles et al., 2013) where we showed that when performing a task on masked stimuli, the Error-related Negativity (ERN), a known brain marker of error detection is present only on conscious trial. In the present study, we replicated this finding using multivariate analysis, showing that the accuracy of motor decisions can be decoded only in conscious conditions. Crucially we now show the presence, in brain activity, of an intention signal which is modulated in exactly the same fashion by subjective visibility and which might serve as an input to error detection and the triggering of the ERN.

The presence of an accurate intention signal independent from the action itself, but contemporaneous with it, readily explains how errors can be detected and corrected, sometimes nearly instantaneously after the wrong key-press (Rabbitt, 2002), or why the ERN starts nearly simultaneously with the erroneous response itself (Rodríguez-fornells et al., 2002). Indeed, this existence of such a signal had been postulated in several previous models which proposed that error-detection results from a comparison (Bernstein et al., 1995; Coles et al., 2001; Maier et al., 2008) or a conflict (Veen and Carter, 2002; Yeung et al., 2004) between the executed and the required response.

A dissociation between intention and action was previously reported by Desmurget and colleagues who found that, during intracranial stimulation of the right inferior parietal region, subjects reported a strong intention to move, without any actual electro-myographic activity (Desmurget et al., 2009; Desmurget and Sirigu, 2012). Similarly, decisions can be decoded from the activity of prefrontal cortex prior to any motor preparation (Haynes et al., 2007). Our finding provides further evidence of a brain representation distinct from the ongoing motor plan, that nonetheless carries information about the intended action. Our decoding method, operating on sensor-level data, did not allow us to investigate directly which brain regions carried this intention signal. However, previous findings suggest that premotor cortex (Gallivan et al., 2011), precuneus (Soon et al., 2013), medial prefrontal cortex (Haynes et al., 2007) and parietal cortex (Desmurget et al., 2009) could be plausible candidates for decoding intentions. In the present experiment, since decoding the required response coincided with decoding the result of the number comparison task, regions involved in number processing (Dehaene et al., 2003) might also be involved. Further research, dissociating these factors, will be needed to specify the precise source of the intention signal that serves as a basis for error detection.

Which computational models may explain how the same system produces an initial error and its subsequent correction? According to some models of decision, a single decision system accounts for both the initial incorrect response and the subsequent corrective action (Kiani and Shadlen, 2009; Resulaj et al., 2009; Pleskac and Busemeyer, 2010) which corresponds to a late "change of mind" (Resulaj et al., 2009). Similarly, connectionist models of conflict hypothesize that, within a single decision network, the wrong decision unit can sometimes be activated in high conflict trials, immediately followed by a rapid correction, triggering a conflict signal reflected by ERN (Botvinick et al., 2001; Yeung et al., 2004). However, these single-representation models are challenged by our findings, which demonstrate the simultaneous presence of two orthogonal patterns of brain activity coding respectively for the ongoing and the required response, and suggest that neural codes for the desired response and the executed motor response are not activated sequentially but in parallel.

This parallel activation fits better with a dual-route model for conscious versus non-conscious processes (Del Cul et al., 2009) in which intentions emerge from the computation of a slow and accurate conscious route. We previously suggested that the dual-route model could account for our observation of an all-or-none error detection reflected by the ERN and triggered only in conscious trials (Charles et al., 2013). Indeed the ERN and its following positive component, the Pe (Falkenstein et al., 2000), as well as patterns of brain activity originating from cingulate cortex (Debener et al., 2005; Charles et al., 2013) could have been at the origin of the signals used by the present performance accuracy decoder. Another related prediction of the model is that the size of the discrepancy (Scheffers and Coles, 2000) or conflict (Steinhauser and Yeung, 2010) between intended and executed action should predict the size of the internal error signal. Indeed, we found that on conscious trials, the trial-by-trial product of action and intention decoding scores correlated with the decoded probability of

accuracy. Likewise, several studies found that the ERN and the Pe vary with the objective amount of evidence in favor of the correct response (Hughes and Yeung, 2011; Steinhauser and Yeung, 2012; Charles et al., 2013) as well as the subjective identification of the required response (Scheffers and Coles, 2000; O'Connell et al., 2007; Dhar et al., 2011; Hughes and Yeung, 2011; Wessel et al., 2011; Shalgi and Deouell, 2012). Furthermore, we found that the time at which this accuracy code emerged correlated with the slowest of the action and intention codes, in accordance with the prediction that the latency of error detection should reflect the latest of the two available signals for action and intention (Van Veen and Carter, 2002; Yeung et al., 2004). While more detailed investigations will be needed to understand the exact determinants of the amplitude and timing of the ERN and the Pe, our findings are in accordance with models that view these components as essential steps of the error detection process (Steinhauser and Yeung, 2010; Wessel et al., 2011; Wessel, 2012).

The present study also sheds light on the distinction between subliminal and conscious processing. We found a dissociation between early and late stages of stimulus processing, consistent with the findings that automatized perceptual, cognitive and motor operations are preserved even under subliminal conditions (Del Cul et al., 2006; Melloni et al., 2007) while later stages show an all-or-none dissociation between conscious and non-conscious trials (Sergent and Dehaene, 2004; Del Cul et al., 2007). Nonetheless, cognitive processes related to performance monitoring may also be partially triggered non-consciously (Nieuwenhuis et al., 2001; Cohen et al., 2009; Logan and Crump, 2010). Indeed, in our previous analysis of the present dataset, we found that subjects could detect the accuracy of motor decisions with above-chance accuracy even on unseen trials (Charles et al., 2013), suggesting that some performance evaluation processes distinct from the ERN operate non-consciously. According to the dual-route model, the level of evidence reached by the non-conscious route is a noisy indicator of the confidence in the response (Galvin et al., 2003; Pleskac and Busemeyer,

2010), and thus may be used as a subliminal index of accuracy. Crucially however, this mechanism is only statistical in nature, and thus unable to confidently and categorically label a given trial as correct as erroneous. Our results suggest that such categorical meta-cognitive knowledge cannot be attained unconsciously, but requires an explicit representation of the required action.

References

Bernstein PS, Scheffers MK, Coles MGH (1995) “Where did I go wrong?” A psychophysiological analysis of error detection. J Exp Psychol Hum Percept Perform 21:1312–1322.

Bode S, Haynes J-D (2009) Decoding sequential stages of task preparation in the human brain. NeuroImage 45:606–613.

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

Chang C, Lin C (2011) LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology:1–39.

Charles L, Van Opstal F, Marti S, Dehaene S (2013) Distinct brain mechanisms for conscious versus subliminal error detection. NeuroImage 73:80–94.

Cohen MX, Van Gaal S, Ridderinkhof KR, Lamme V a F (2009) Unconscious errors enhance prefrontal-occipital oscillatory synchrony. Frontiers in human 3:1–12.

Coles MGHM, Scheffers MMK, Holroyd CCB (2001) Why is there an ERN/Ne on correct trials? Response representations, stimulus-related components, and the theory of error-processing. Biological psychology 56:173–189.

Dehaene S, Piazza M, Pinel P, Cohen L (2003) Three parietal circuits for number processing. Cognitive neuropsychology 20:487–506.

Del Cul A, Baillet S, Dehaene S (2007) Brain dynamics underlying the nonlinear threshold for access to consciousness. PLoS Biol 5:2408–2423.

Del Cul A, Dehaene S, Reyes P, Bravo E, Slachevsky A (2009) Causal role of prefrontal cortex in the threshold for access to consciousness. Brain 132:2531–2540.

Del Cul A Del, Dehaene S, Leboyer M, Del Cul A (2006) Preserved subliminal processing and impaired conscious access in schizophrenia. Archives of general psychiatry 63:1313.

Desmurget M, Reilly KT, Richard N, Szathmari A, Mottolese C, Sirigu A (2009) Movement intention after parietal cortex stimulation in humans. Science 324:811–813.

Desmurget M, Sirigu A (2012) Conscious motor intention emerges in the inferior parietal lobule. Current opinion in neurobiology 22:1004–1011.

Dhar M, Wiersema JR, Pourtois G (2011) Cascade of Neural Events Leading from Error Commission to Subsequent Awareness Revealed Using EEG Source Imaging Sirigu A, ed. PLoS ONE 6:e19578.

Gallivan JP, McLean DA, Valyear KF, Pettypiece CE, Culham JC (2011) Decoding action intentions from preparatory brain activity in human parieto-frontal networks. The Journal of neuroscience 31:9599–9610.

Galvin SJ, Podd J V, Drga V, Whitmore J (2003) Type 2 tasks in the theory of signal detectability: discrimination between correct and incorrect decisions. Psychonomic bulletin & review 10:843–876.

Gehring WJ, Goss B, Coles MGH, Meyer DE, Donchin E (1993) A neural system for error detection and compensation. Psychological Science 4:385–390.

Haynes J-D, Sakai K, Rees G, Gilbert S, Frith C, Passingham RE (2007) Reading hidden intentions in the human brain. Current biology 17:323–328.

Holland PW, Welsch RE (1977) Robust regression using iteratively reweighted least-squares. Communications in Statistics - Theory and Methods 6:813–827.

Hughes G, Yeung N (2011) Dissociable correlates of response conflict and error awareness in error-related brain activity. Neuropsychologia 49:405–415.

Kiani R, Shadlen MN (2009) Representation of confidence associated with a decision by neurons in the parietal cortex. Science 324:759–764.

Lemm S, Blankertz B, Dickhaus T, Müller K-R (2011) Introduction to machine learning for brain imaging. NeuroImage 56:387–399.

Logan GD, Crump MJC (2010) Cognitive illusions of authorship reveal hierarchical error detection in skilled typists. Science 330:683.

Maier ME, Steinhauser M, Hubner R (2008) Is the error-related negativity amplitude related to error detectability? Evidence from effects of different error types. Journal of Cognitive Neuroscience 20:2263–2273.

Maris E, Oostenveld R (2007) Nonparametric statistical testing of EEG- and MEG-data. Journal of neuroscience methods 164:177–190.

Melloni L, Molina C, Pena M, Torres D, Singer W, Rodriguez E (2007) Synchronization of neural activity across cortical areas correlates with conscious perception. The Journal of neuroscience 27:2858–2865.

Nieuwenhuis S, Ridderinkhof KR, Blom JH, Band GPH, Kok A (2001) Error-related brain potentials are differentially related to awareness of response errors: evidence from an antisaccade task. Psychophysiology 38:752–760.

Norman D, Shallice T (1986) Attention to action: Willed and automatic control of behavior. In: R. J. Davidson, G. E. Schwarts, & D. Shapiro (Eds.). Consciousness of self-regulation. Advances in research and theory (Vol. 4). New York: Plenum Press, pp 1–18.

Norman DA (1981) Categorization of Action Slips. Psychological Review 88:1–15.

Norman K a, Polyn SM, Detre GJ, Haxby J V (2006) Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends in cognitive sciences 10:424–430.

O'Connell RG, Dockree PM, Bellgrove MA, Kelly SP, Hester R, Garavan H, Robertson IH, Foxe JJ (2007) The role of cingulate cortex in the detection of errors with and without awareness: a high-density electrical mapping study. Eur J Neurosci 25:2571–2579.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É (2011) Scikit-learn: Machine Learning in Python. The Journal of Machine Learning Research 12:2825–2830.

Platt JC (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers 10:61–74.

Pleskac TJ, Busemeyer JR (2010) Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. Psychological review 117:864–901.

Posner MI, Rothbart MK (1998) Attention, self-regulation and consciousness. Philosophical transactions of the Royal Society of London Series B, Biological sciences 353:1915–1927.

Rabbitt PMA (1966) Error correction time without external error signals. Nature 212:438.

Rabbitt PMA (2002) Consciousness is slower than you think. Q J Exp Psychol A 55:1081–1092.

Resulaj A, Kiani R, Wolpert DM, Shadlen MN (2009) Changes of mind in decision-making. Nature 461:263–266.

Rodríguez-fornells A, Kurzbuch AR, Münte TF (2002) Time course of error detection and correction in humans: neurophysiological evidence. The Journal of neuroscience 22:9990–9996.

Scheffers MK, Coles MGH (2000) Performance monitoring in a confusing world: error-related brain activity, judgments of response accuracy, and types of errors. J Exp Psychol Hum Percept Perform 26:141–151.

Sergent C, Dehaene S (2004) Neural processes underlying conscious perception: experimental findings and a global neuronal workspace framework. J Physiol Paris 98:374–384.

Shalgi S, Deouell LY (2012) Is any awareness necessary for an Ne? Frontiers in human neuroscience 6:1–15.

Soon CS, He a. H, Bode S, Haynes J-D (2013) Predicting free choices for abstract intentions. Proceedings of the National Academy of Sciences.

Steinhauser M, Yeung N (2010) Decision processes in human performance monitoring. The Journal of Neuroscience 30:15643–15653.

Steinhauser M, Yeung N (2012) Error awareness as evidence accumulation: effects of speed-accuracy trade-off on error signaling. Frontiers in human neuroscience 6:240.

Van Veen V, Carter CS (2002) The timing of action-monitoring processes in the anterior cingulate cortex. Journal of cognitive neuroscience 14:593–602.

Veen V Van, Carter CS (2002) The anterior cingulate as a conflict monitor: fMRI and ERP studies. Physiol Behav 77:477–482.

Wessel JR (2012) Error awareness and the error-related negativity: evaluating the first decade of evidence. Frontiers in human neuroscience 6:88.

Wessel JR, Danielmeier C, Ullsperger M (2011) Error awareness revisited: accumulation of multimodal evidence from central and autonomic nervous systems. Journal of Cognitive Neuroscience 23:3021–3036.

Yeung N, Botvinick MM, Cohen JD (2004) The Neural Basis of Error Detection: Conflict Monitoring and the Error-Related Negativity. Psychological Review 111:931–959.

Tables

Table 1: Statistical results of within-subject classification scores when decoding the required response on *Seen* trials, separately for left or right motor response.

	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Subject 6
Actual Response Left	n.s.	**	n.s.	***	***	***
Actual Response Right	***	~	~	***	***	***

Note: n.s. $p>0.1$, ~$p<0.1$, *$p<0.05$, **$p<0.01$, ***$p<0.001$

Figures

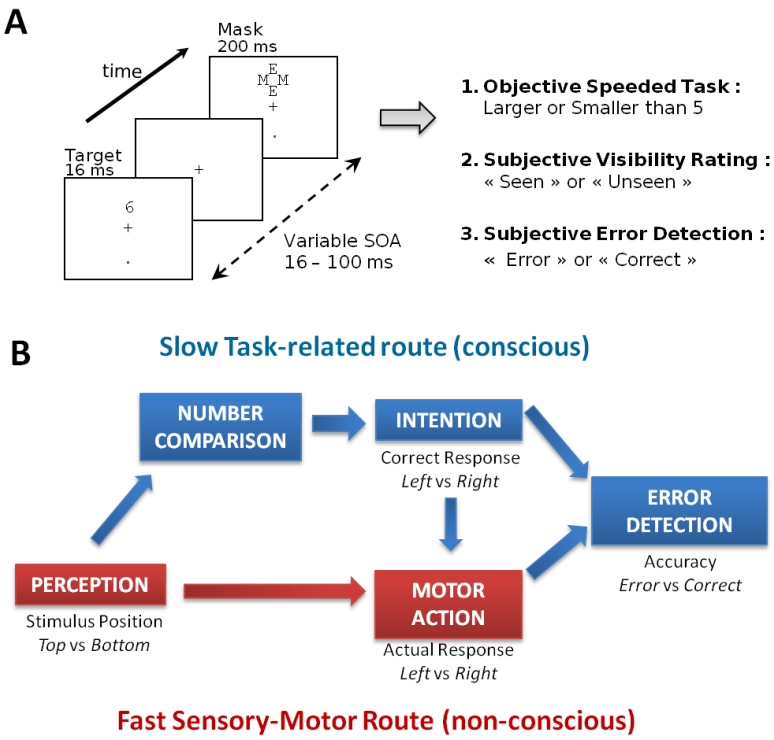


Figure 1 : Experimental Design and Dual-Route Model.

(A) On each trial, a number was presented for 16 ms at one of two possible locations (top or bottom). It was followed by a mask composed of a fixed array of letters presented at a varying duration after target onset (16, 33, 50, 66 or 100 ms). Participants first performed a speeded forced-choice number comparison task where they decided whether the number was smaller or larger than 5. Then, they evaluated the subjective visibility of the target and their own performance in the primary number comparison task.

(B) Dual-Route Model for error-detection. In this model, two routes accumulate sensory evidence in parallel. A response is emitted by whichever route first reaches its decision threshold. The first route corresponds to automatic sensory-motor association and can be triggered non-consciously to produce fast motor responses. The second-route corresponds to the slower, voluntary processing of the stimulus according to task instructions and produces a conscious representation of the required response, i.e. a conscious intention. The comparison of the outputs of these two routes allows participants to detect a discrepancy between their intended and ongoing responses, and therefore to self-evaluate their performance.

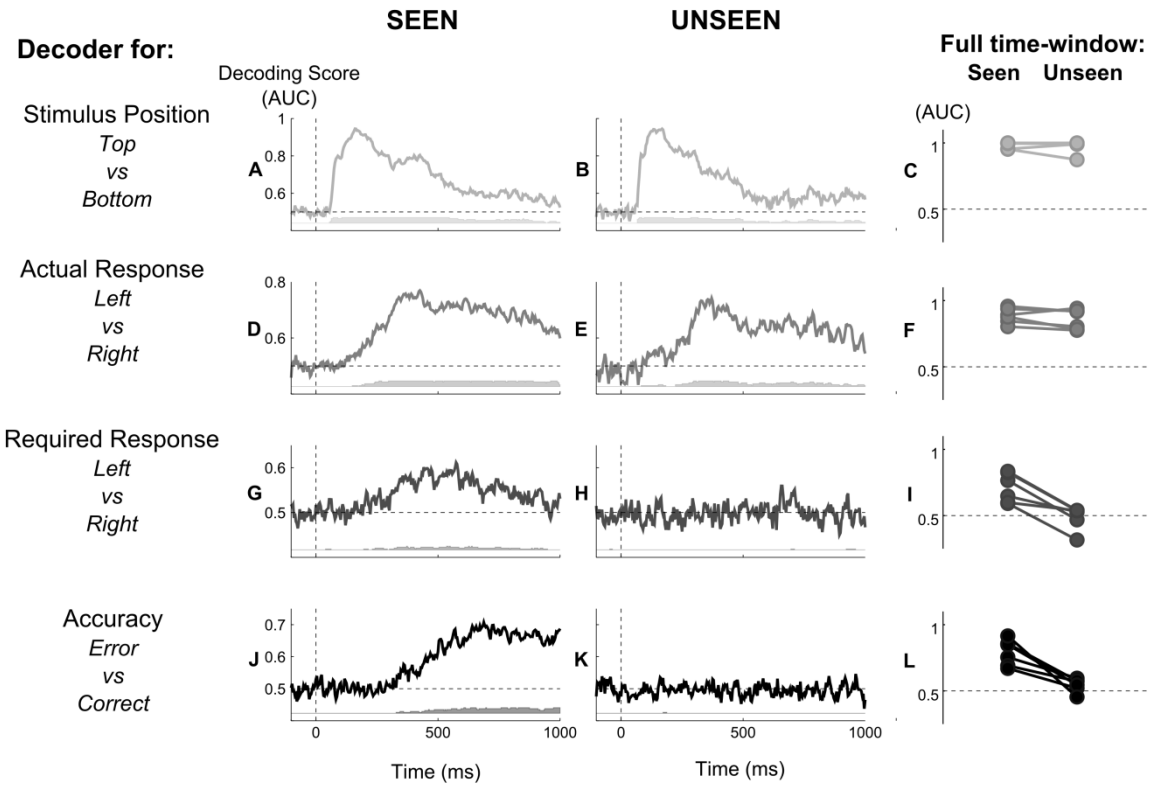


Figure 2: Decoding Perception, Action, Intention and Accuracy, for conscious and non-conscious trials.

Multivariate decoding was applied either to each time sample (central columns) or to the full trial time-window (right columns). Results demonstrate that while Stimulus Position and Actual Response could be decoded in both conscious and non-conditions with high accuracy, the Required Response and the Accuracy could be decoded solely in conscious conditions.

(Central columns) Area under the curve (AUC), a measure of decoding accuracy, is plotted after averaging across subjects, aligned on stimulus onset, separately for the stimulus position decoder (top versus bottom, A-B), actual response decoder (left versus right, D-E), required Response decoder (left versus right, G-H) and accuracy decoder (error versus correct, J-K) respectively in *seen* (A,C,E,G) and *unseen* (B,D,F,H) conditions. Gray bars below each graph indicate, for each time-point, the number of subjects presenting an above-chance classification score at that instant as computed by cluster analysis.

(Right column) For each subject, individual measures of AUC are plotted for *Seen* (left) and *Unseen* (right) conditions, separately for the Stimulus Position decoder (C), Actual response decoder (F), Required Response decoder (I) and Accuracy decoder (L). In each case, decoding was applied on all the sensors and time points from the full trial time-window (0-800 ms after stimulus presentation).

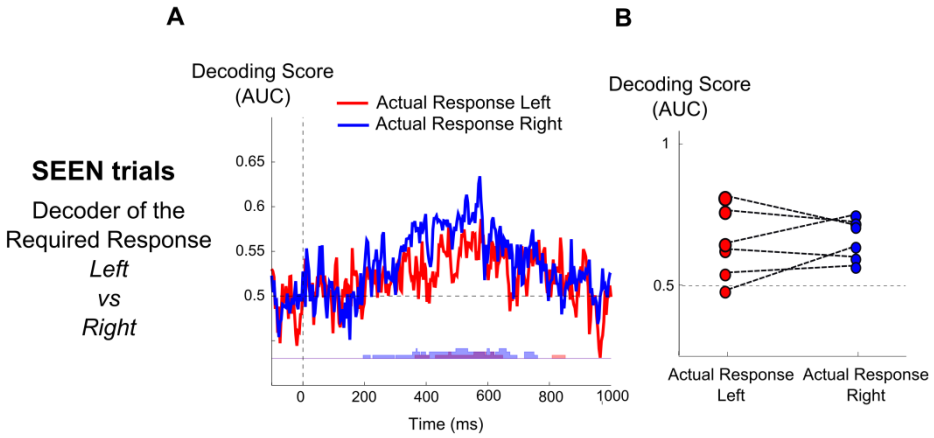


Figure 3: Decoding conscious intention independently of motor action

This figure demonstrates that, while subjects are preparing for a given response (correct or erroneous) their brain activity contains decodable information about the response that they *should* make (the required response).

(A) Average measure of AUC across subjects when decoding the Required Response on *Seen* trials, separately for Left (red line) or Right (blue line) actual motor responses. Time zero corresponds to the onset of the stimulus. Bars below graph indicate for each time-point the number of subjects presenting an above-chance classification score at that instant as computed by cluster-analysis.

(B) For each subject, individual AUC measures when decoding the required response on *Seen* trials, separately for Left (red points) or Right (blue points).

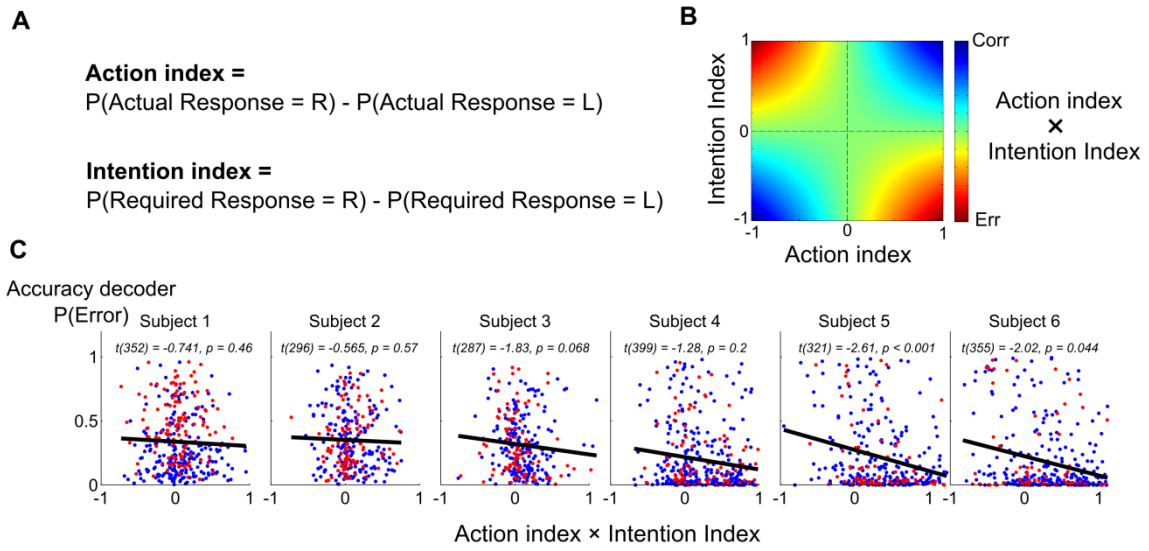


Figure 4: Congruity between action and intention correlates with the strength of error decoding.

(A) To obtain a trial-by-trial measure of the strength of internal representations of action and intention, we first transformed the output of the classifiers by subtracting the classification probability of the Left Response from the classification probability of the Right response, thus yielding for each trial a measure ranging from -1 (i.e. certainly of a left response) to +1 (certainty of a Right Response). This computation was done separately for the actual response and for the required response, thus yielding two single-trial indices of the strength of internal representations, the action index and the intention index.

(B) The product of the Intention and Action indices reflects the congruity between intended and executed actions. Positive values (Blue) are obtained when both action and intention are congruent (the values of the two indices are of the same sign), indicating a high probability of being correct. On the opposite, negative values (Red) indicate a discrepancy between action and intention, and therefore a high probability of committing an error. Note than when no information is available on either the action or the intention, the product is close to 0 and does not allow distinguishing error from correct trials

(C) Correlation results of the product of Action and Intention indices with the decoded Error Probability for each subject. Each dot corresponds to a single *seen* trial (red = errors, blue = correct). A negative correlation confirms that the internal representation of an upcoming error is stronger when the discrepancy between internal representations of action and intention is larger.

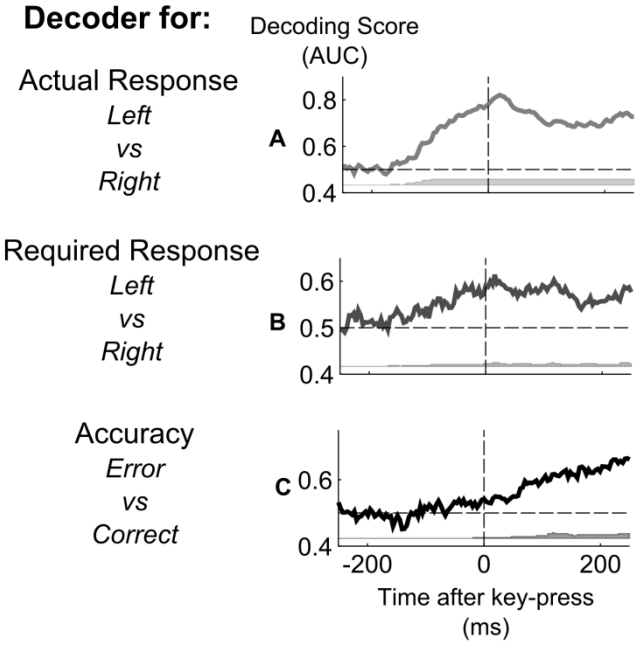


Figure 5: Decoding Action, Intention and Accuracy before and after the actual key press.

For *Seen* trials only, the figure shows the time course of decoding the Actual Response (A), the Required Response (B) and the Accuracy (C), relative to actual key press. The curves were realigned on motor onset and an average measure of decoding success (area under the curve, AUC) was computed across subjects. Gray bars below graph indicate for each time-point the number of subjects presenting an above-chance classification score at that instant as computed by cluster-analysis.

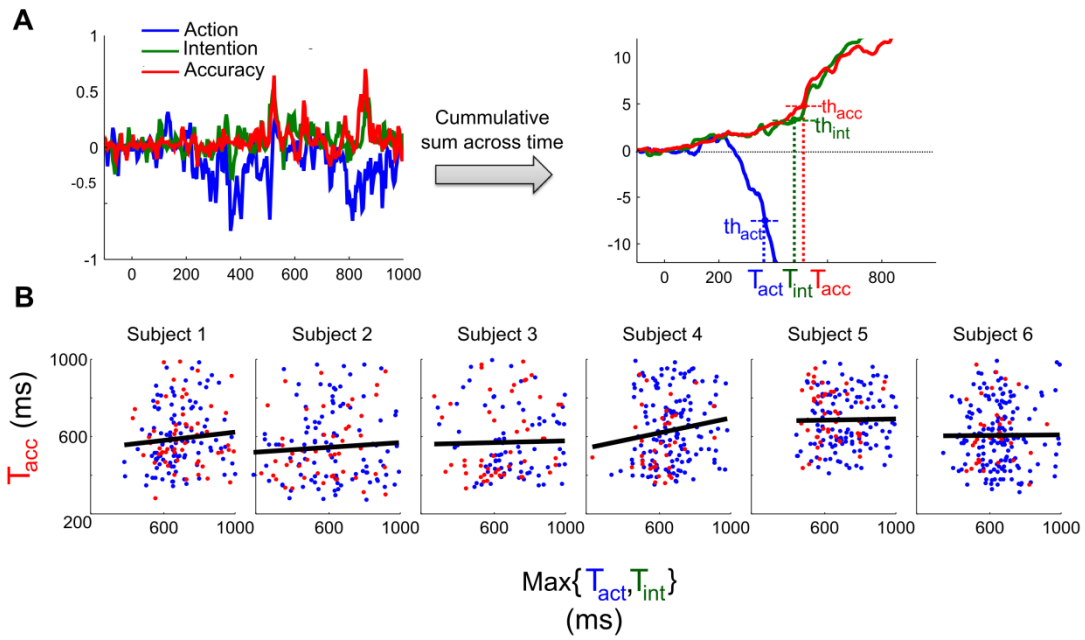


Figure 6: The timing of error detection correlates with the slowest of two signals for action and intention

(A) Example of a single-trial computation of decoding time. To improve the signal-to-noise ratio, we computed the cumulative sum, across time, of the probability values obtained from each of the three decoders for actual response, required response, and accuracy. Threshold values for each decoder were defined, and the timing of threshold crossing for each time-series was taken as an index of the time when this code first became available on this trial. Thus, three values were obtained for each trial : T_{int} ,T_{act} and T_{acc} corresponding respectively to the time for threshold crossing of the actual response, the required response and the accuracy decoder

(B) Correlation results of the slowest (maximum) time index between T_{int} and T_{act} with the time index of error detection T_{acc}. Each dot corresponds to a single *seen* trial (red = errors, blue = correct). A positive correlation indicates that, as predicted, error information becomes available only once both action and intention codes have been computed.

CHAPTER 6

Article 3 : Preserved unconscious metacognition and impaired conscious error-detection in schizophrenia

---

6.1 Introduction to the article

6.1.1 Context and goal of the study

In our initial study (Charles et al., 2013), we showed that the ERN is triggered only in conscious conditions, when subjects reported consciously seeing the target. We were also able to demonstrate the existence, in addition to the ERN, of another active system in non-conscious conditions that allows subjects to predict their level of performance. We suggested that distinct metacognitive mechanisms may be involved in conscious and non-conscious conditions: all-or-none error detection, indexed by the ERN is present only in conscious conditions, but confidence in one's response can still be computed under non-conscious conditions.

Schizophrenia has been associated with specific deficits in conscious processing but not in non-conscious processing (Tononi and Edelman, 2000; Danion et al., 2001). Indeed, in one study Dehaene et al. (2003) demonstrated that in a task causing a conflict between two contradictory responses, patients had decreased brain markers related to conflict monitoring only for conscious stimuli while classical effects in the subliminal conditions were preserved. A second study (Del Cul et al., 2006) confirmed these results, showing that patients with schizophrenia have a threshold for consciousness that is higher compared to controls while non-conscious processing (measured by subliminal priming) of the visual stimuli are preserved.

Our paradigm allowed us to disentangle conscious and non-conscious mechanisms linked to performance monitoring and therefore to determine specific deficits associated with conscious functions. Indeed, the fact that the ERN was conditioned by conscious perception allows us to make the prediction that any disturbance in conscious process should induce alteration of the ERN. However, the non-conscious brain processes of error detection that are distinct from those at stake in the conscious condition should not be altered. Therefore, such a paradigm is particularly relevant for schizophrenic patients, allowing us to understand how high-order cognitive functions linked to metacognition and conscious monitoring are affected in these types of pathology. Furthermore, it constitutes a test case for our

model that distinguishes metacognitive processes in conscious and non-conscious conditions.

To determine whether conscious and non-conscious metacognitive processes truly dissociate and if schizophrenic patients present a specific alteration of these processes according to conscious access, we replicated our initial experiment in a population of schizophrenic patients. The protocol was similar to the one used in the second experiment of our first study (Charles et al., 2013). We tested thirteen schizophrenic patients and thirteen age-matched healthy controls to determine how metacognitive processes were affected in schizophrenia and how they were modulated by consciousness

6.1.2 Summary of the results

We found that schizophrenic patients presented altered responses to error in conscious conditions manifested by a decreased ERN. However non-conscious process linked to computing the likelihood of having made an error were not altered for these patients.

Interestingly, in *seen* trials for which the target number was consciously perceived, both first-order and second-order performance were significantly lower in schizophrenic patients compared to control subjects. Indeed, while committing more errors than control subjects, schizophrenic patients were also able to detect a lower fraction of their errors, a result that is in line with previous studies showing impaired metacognitive ability in schizophrenia (Bates et al., 2002). Importantly, the reduced performance in error detection was associated with reduced amplitude for the ERN compared to control subjects. Such results fit with previous findings in electrophysiological studies of brain responses to errors in schizophrenic patients showing a reduced ERN in schizophrenia (Kerns et al., 2005; Foti et al., 2012; Bates et al., 2004; Bates et al., 2002; Mathalon, 2002; Kopp and Rist, 1999; Morris et al., 2006; Morris et al., 2011; Alain, 2002; Kim et al., 2006; Olvet and Hajcak, 2008; Carter et al., 2001; Laurens, 2003; Hajcak et al., 2004; Pailing and Segalowitz, 2004b). This finding was also coherent with results of MEG studies, which presented distinct patterns of activity in conscious trials for patients compared with controls. Our results on control subjects replicated our previous study showing that both PCC and ACC are strongly active in error compared to correct trials, with a peak of activity simultaneous to the ERN maximum. For schizophrenic patients however, only small patches of the cingulate cortex were found to be active, coherent with the reduced ERN amplitude.

In *unseen* trials however, we found that schizophrenic patients were able to perform the number-comparison task slightly better than chance, with a similar level of accuracy as controls. More importantly, they were also able to evaluate the accuracy of their own decision in unseen trials as well as control subjects. This finding, which replicates our previous results (Charles et al., 2013) and extends them to schizophrenic patients suggest that non-conscious processes of evaluation of the confidence in one's response can operate non-consciously. More importantly, it suggests that these processes are preserved in schizophrenia.

Crucially, this finding confirms our initial hypothesis that these non-conscious statistical evaluation processes are distinct from that reflected by the ERN. We found previously that different patterns of ac-

tivity were evoked by the ERN and by non-conscious metacognitive processes, suggesting an anatomical distinction between the two (Charles et al., 2013b). The present results confirm this result, showing that conscious mechanisms of error-detection reflected in the ERN can be impaired whilst non-conscious performance monitoring processes are preserved.

6.2 Article

Charles, L. & Dehaene, S. 2013 Preserved unconscious metacognition and impaired conscious error-detection in schizophrenia. In preparation

Preserved unconscious metacognition and impaired conscious error-detection in schizophrenia

Lucie Charles [a, b, c]

Raphaël Gaillard [e, f, g]

Isabelle Amado [h]

Stanislas Dehaene [a, b, c, d]

[a] INSERM, U992, Cognitive Neuroimaging Unit,

[b] CEA, DSV/I2BM, NeuroSpin Center

[c] Univ Paris-Sud, Cognitive Neuroimaging Unit,

[d] Collège de France,

[e] INSERM U894, Centre de Psychiatrie & Neurosciences, Paris, France;

[f] Université Paris Descartes, Sorbonne Paris Cité, Paris, France;

[g] Centre hospitalier Sainte-Anne, Service Hospitalo-Universitaire, Paris, France

[h] CH Sainte-Anne, C3RP & Services de Psychiatrie d'adultes HU & 75G17, Paris, France

Corresponding author: Lucie Charles

INSERM-CEA Cognitive Neuroimaging unit
CEA/SAC/DSV/DRM/NeuroSpin
Bât 145, Point Courrier 156
F-91191 Gif/Yvette, FRANCE
Tel: +33 1 69 08 99 74
Fax: +33 1 69 08 79 73
lucie.charles.ens@googlemail.com

ABSTRACT

The ability to detect our own errors constitutes a crucial function of action monitoring processes. In a recent study, we found that some performance monitoring processes could occur outside of awareness, while others were tightly linked to conscious perception. In particular, the ERN, a known brain marker of performance monitoring was present solely in conscious perception. However, even in the absence of an ERN, subjects could still evaluate better than chance the confidence in their response in non-conscious trials. This result suggests that distinct brain processes related to performance monitoring are triggered in conscious and non-conscious conditions.

Schizophrenia have been associated with altered conscious access to mental content while non-conscious processes such as priming remain unimpaired. Indeed, the ERN was found to be drastically reduced in schizophrenic patients. To verify that error detection processes in conscious and non-conscious conditions were computationnaly distinct, we replicated our initial experiment in a population of schizophrenic patients. Thirteen patients with schizophrenia and thirteen control subjects matched in age performed a speeded number comparison task on masked stimulis. Importantly, conscious perception and error-detection were assessed on a trial-by-trial basis by mean of subjective report of visibility and confidence. We found that schizophrenic patients presented altered responses to error in conscious conditions which manifested by a decreased ERN. However non-conscious process linked to computing the likelihood of having made an error were not altered for these patients. These results confirm the dissociation of conscious and non-conscious metacognitive processes, suggesting that deficit in schizophrenia are specifically linked to conscious processing while non-conscious processes remain unimpaired.

I. INTRODUCTION

Monitoring of performance is a crucial feature for cognitive control. In particular, error-detection constitutes a key element for adaptive behaviour, allowing to adjust attention and top-down control resources to avoid further mistakes. In a recent study in healthy subjects (Charles et al., 2013b), we found that distinct performance monitoring processes could be triggered in conscious and non-conscious conditions. In particular, we showed that the Error-related Negativity (ERN), a know brain marker linked to error detection was triggered only in conscious conditions, while above-chance estimation of performance was still possible outside of awareness. This result suggested that even in non-conscious trials some processes related to the evaluation of the confidence in the response could be triggered, even in the absence of an ERN signal. We proposed that these results could be explained by the existence of two distinct systems for performance monitoring: a conscious signal based on the ERN and reflecting the all-or-none detection of an error and a non-conscious process estimating in a statistical manner the confidence in the response.

Patients suffering from schizophrenia have been shown to present deficits in these executive functions and in particular error detection. Indeed, the ERN has been reported to be abnormally reduced in schizophrenic patients (Kopp and Rist, 1999; Alain, 2002; Bates et al., 2002, 2004; Mathalon, 2002; Kim et al., 2006; Morris et al., 2006; Foti et al., 2012). This result is consistent with other findings showing abnormal activation in schizophrenic patients of prefrontal regions (Barch et al., 2001) and in particular in cingulate cortex (Carter et al., 2001; Dehaene et al., 2003; Laurens, 2003a; Kerns et al., 2005; Polli et al., 2008; Yan et al., 2012) linked to abnormal inhibition (Hughes et al., 2011), conflict monitoring (Dehaene et al., 2003; Kerns et al., 2005) or error detection processes (Bates et al., 2002, 2004).

What are the origins of these deficits? Altered cognitive functions in schizophrenia have been associated with abnormalities of connectivity patterns among distant brain regions (Friston and Frith, 1995; Friston, 1998, 2005; Haraldsson, 2004; Liang et al., 2006; Uhlhaas and Singer, 2010; Schmitt et al., 2011), in particular in prefrontal cortex (Fletcher et al., 1999; Grillon et al., 2012). Interestingly, such abnormalities could also explain altered conscious access in schizophrenic patients (Tononi and Edelman, 2000; Dehaene et al., 2003; Del Cul et al., 2006), in the line of theories of consciousness relying on connectivity between distant brain areas as a key feature for conscious processes. According to Integrated Information Theory, impaired connectivity would be associated a deficit in integration of information that would strongly impact conscious processes (Tononi and Edelman, 2000). Similarly, Global Workspace theory hypothesizes that conscious access is tightly linked to long-range connections between remote brain areas (Dehaene and Changeux, 2011), any disruption in these connection leading to impaired conscious access. Indeed, specific deficits in conscious versus non-conscious conditions have been reported in several studies on schizophrenic patient. While non-conscious processes such as priming (Dehaene et al., 2003), implicit learning (Danion et al., 2001) or task inhibition (Huddy et al., 2009) appear to be preserved in schizophrenic patients, conscious processing are characterized by reduced activity and impaired cognitive functions.

Our results (Charles et al., 2013b) could provide a framework to interpret these different findings. By showing that some cognitive control mechanisms, such as the one reflected by the ERN, are specifically associated with conscious access, we could explain how precise executive functions are altered in schizophrenia. Following this hypothesis, data from schizophrenic patients could provide a test-case to determine whether a specific process involves mainly conscious or non-conscious processing.

In order to validate this hypothesis, we replicated our experiment (Charles et al., 2013b) in a population of schizophrenic patients. Our line of reasoning was the following: if conscious and non-conscious performance monitoring processes are truly distinct, patients presenting a specific deficit in conscious conditions should also present an alteration of conscious error-detection in the form of a reduced ERN. On the contrary, non-conscious "meta-performance" corresponding to above-chance accuracy judgment should be preserved for these patients. We tested thirteen schizophrenic patients and thirteen controls subjects matched for age and sex in a similar paradigm. Participants performed a number comparison task on a masked digit, while perceptual evidence was systematically manipulated by varying the target-mask Stimulus Onset Asynchrony (SOA). Crucially, subjective perception was assessed on a trial by trial basis by asking participants to report their visibility of the target (Seen/Unseen) as well as their perceived performance (Error/Correct) in the number comparison task. This approach allowed us to study how the ERN and error-detection performance were modulated by subjective perception of the stimulus (subliminal/subjectively unseen trials versus conscious/seen trials) for schizophrenic patients compared to controls.

II. MATERIALS & METHODS

1.1. Participants.

Twenty-one schizophrenic patients were tested in this experiment (5 women and 16 men; mean age 30.9 years old). Patients met DSM-IV criteria for schizophrenia and were recruited from the psychiatric department of Saint-Anne Hospital (Assistance Publique, Hôpitaux de Paris). They had a chronic course and were stable at the time of the experiment. 17 patients were medicated by atypical antipsychotics, 1 with typical antipsychotics and two were not medicated at the time of the experiment. This treatment had been unchanged for at least three weeks. A trained clinician evaluated and categorized the symptoms of the schizophrenic participants on the day of participation using Positive and Negative Syndrome Scale (PANSS).

One patient was excluded of the protocol due to the discovery on the MRI scan of a prefrontal lesion. Two patients misunderstood the task instructions on the evaluation of decision accuracy and two others reported either never seeing the target number or on the contrary seeing the target even when it was absent in 60% of the trials. These patients were therefore excluded from the analysis. Additionally, three other patients had an insufficient numbers of conscious errors to be included in the analysis. Therefore thirteen patients could be kept for the entire analysis (four women and nine men, mean age 28,8 years old, two left handed).

The comparison group consisted of 13 subjects (mean age, 28,8, range four women and nine men, two left handed). Comparison subjects were excluded for history of any psychotic disorder, bipolar disorder, schizotypal or paranoid personality disorder, recurrent depression. Patients and controls with a history of brain injury, epilepsy, alcohol or substance abuse, or other neurological or ophthalmologic disorders were excluded. Patients and controls did not differ significantly in sex and age. All experiments were approved by the French regional

ethical committee for biomedical research, and subjects gave written informed consent. All participants had normal or corrected-to-normal vision.

1.2. Design & Procedure

A masking paradigm similar to experiment 2 in Charles *et al.* (Charles et al., 2013b) was used in the present study. The target-stimuli (the digits 1, 4, 6, or 9) were presented on a white background screen using E-Prime software. The trial started with a small increase in the size of the fixation cross (100 ms duration) signalling the beginning of the trial. Then the target stimulus appeared for 16 ms at one of two positions (top or bottom, 2.29 degrees from fixation), with a 50 % probability. After a variable delay, a mask appeared at the target location for 250ms. The mask was composed of four letters (two E's and two M's, see Figure 1) tightly surrounding the target stimulus without superimposing or touching it. The stimulus-onset asynchrony (SOA) between the onset of the target and the onset of the mask was varied across trials. Five SOAs were randomly intermixed: 16, 33, 50, 66 and 100ms. The foreperiod duration was manipulated so that the mask always appeared 800 ms after the signal of the beginning of the trial. In one sixth of the trials, the target number was replaced by a blank screen with the same duration of 16ms (mask-only condition), allowing us to study visibility ratings when no target was presented.

Participants primarily had to perform a forced-choice task of comparing the target number to the number 5. Responses were collected 2000 ms after target onset with two buttons using the index of each hand (left button press = smaller-than-5; right button-press = larger-than-5 response). To induce errors, participants were instructed to respond as fast as they could just after the appearance of the target.

At the end of each trial, after another delay of 500 ms, participants were requested to provide two subjective answers with no time-pressure. The first answer was related to the subjective visibility of the target number. In this visibility task, participants had to indicate if

they saw a target number or not. The second answer concerned the participants' knowledge of their performance. Here, they had to indicate whether they thought they had made an error or not in the number comparison task (performance evaluation task). Instructions were clearly stated to ensure that participants understood that the performance evaluation task was directed to the number comparison task and not the visibility judgment. Furthermore, participants were informed that, even when they had not seen the stimulus and thought that they responded randomly, they still had a 50 % chance of having made a correct response. Therefore, they were told to hazard a guess on their performance, even when they did not see the stimulus. For both subjective responses, words corresponding to the two responses (*seen*/*unseen* and *error/correct*) were displayed on the screen and participants had to use the corresponding-side buttons to answer. The words were presented at randomized left and right locations (2.3 degrees from fixation) to ensure that participants didn't use automatized button-press strategy.

The experiment was divided in blocks of 48 trials. Each block contained 8 trials for every SOA condition, with each digit presented at the two possible target locations (Top/Bottom). Participants performed 8 or 11 blocks during EEG/MEG recording. Ten trials of the experiment were given as training before starting the actual recording.

1.3. Simultaneous EEG and MEG recordings.

Simultaneous recording of MEG and EEG data was performed. The MEG system (the Elekta-Neuromag) comprised 306 sensors: 102 Magnetometers and 204 orthogonal planar gradiometers (pairs of sensors measuring the longitudinal and latitudinal derivatives of the magnetic field). The EEG system consisted of a cap of 60 electrodes with reference on the nose and ground on the clavicle bone. Six additional electrodes were used to record electrocardiographic (ECG) and electro-oculographic (vertical and horizontal EOG) signals.

A 3-dimensional Fastrak digitizer (Polhemus, USA) was used to digitize the position of three fiducial head landmarks (Nasion and Pre-auricular points) and four coils used as indicators of

head position in the MEG helmet, for further alignment with MRI data. Sampling rate was set at 1000 Hz with a hardware band-pass filter from 0.1 to 330 Hz.

1.4. SDT analysis

To obtain an unbiased measure of visibility and performance, we used Signal Detection Theory (SDT) to compute *d'*= z(HIT)- z(FA) for the target-detection task (*detection-d'*, where HIT=proportion of trials with target present and response *seen*, and FA= proportion of trials with target absent and response *unseen*) and the number comparison task (where HIT=proportion of trials with target smaller than 5 and a left response, and FA= proportion of trials with target larger than 5 and a left response).

The *meta-d'* measure was computed according to Maniscalco *et al.* (Maniscalco and Lau, 2012). Briefly, classic SDT can be extended to predict what should be the theoretical performance in meta-cognitive judgements where one must evaluate one's own primary performance, such as confidence ratings or error detection. The theory assumes that both primary and meta-cognitive judgements have access to the same stimulus sample on the same continuum. First-order judgments are performed by setting a first criterion in the middle of the continuum. Meta-cognitive judgements are performed by setting two additional criteria surrounding the first-order one, and responding "error" if the sample falls between these two criteria, or "correct" if the sample falls beyond them (i.e. a sample distant enough from the first-order criterion signals high confidence in the primary response). From this ideal-observer theory, precise mathematical relations linking performance and meta-performance can be deduced (Galvin et al., 2003) and it is possible to compute a second-order measure of meta-performance by classifying meta-cognitive responses as second-order hits and false alarm. However, the traditional measure of *d'* does not directly apply to a second-order task because it is not unbiased (second-order *d'* systematically depends on the first-order criterion) and the assumption of normality of the distributions is violated. In order to obtain a valid measure of

meta-performance, unbiased and comparable to the first-order d', Maniscalco *et al.* (http://www.columbia.edu/~bsm2105/type2sdt/) proposed an alternative solution, *meta-d'*. Their proposal consists in bringing both first and second-order performance to the same scale, by determining what should have been the *d'* in the first-order task given the observed second-order (meta) performance, under the assumption that the subject used exactly the same information in both cases. Since meta-d' is expressed in the same scale as d', the two can be compared directly. When *meta-d' < d'*, it means that the subject did worse in the performance evaluation task than expected according to his actual *d'* value. On the opposite, if the *meta-d'* > d', it means that more information was available for subjective performance evaluation than for the primary objective decision.

Meta-d' was estimated by fitting the parameters of a type-I SDT model so that the predicted type-II hits and false-alarm rates were fitted to the actual type-II data. Therefore, meta-d' corresponds to the d' that maximizes the likelihood of the observed type performance, assuming the same bias of response as the one observed in the data.

1.5. MEG/EEG Data Analysis.

MEG data were first processed with MaxFilter™ software using the Signal Space Separation algorithm. Bad MEG channels were detected automatically and manually, and interpolated. Head position information recorded at the beginning of each block was used to realign head position across runs and transform the signal to a standard head position framework.

To remove the remaining noise, Principal Component Analysis (PCA) was used. Artifacts were detected on the electro-occulogram (EOG) and electro-cardiogram. Data were averaged on the onset of each blinks and heart beats separately and PCA was performed

separately for each type of sensor. Then, one to three of the first components characterizing the artifact were selected by mean of visual inspection to be further removed.

Data were then entered into Matlab software and processed with Fieldtrip software (http://fieldtrip.fcdonders.nl/). For each channel, a manual rejection of trials based on signal discontinuities was performed and the discarded trials interpolated from surrounding channels. A low-pass filter at 30 Hz was then applied as well as a baseline correction from 300 ms to 200 ms before target onset.

Data were then realigned on response onset to be further averaged by subject and conditions. To obtain grand-average evoked response data, we first averaged individual data for each SOA separately, then averaged across SOAs and then across participants. A baseline correction was performed from 200 to 50ms before motor response.

1.6. Combined EEG/MEG Source Reconstruction

Brainstorm software was used to derive current estimate from correct and error MEEG waveforms, for each condition of visibility and each subject separately. Cortical surfaces of 23 participants (1 patient and 2 control were discarded as no MRI data could be obtained) were reconstructed from individual MRI with FreeSurfer (http://surfer.nmr.mgh.harvard.edu/). Inner skull and outer-skull surfaces were estimated using an additional flash sequence and MNE software (Fischl et al., 2004), in order to compute accurate forward model using a three-compartment boundary-element method (OpenMeeg toolbox; http://www-sop.inria.fr/athena/software/OpenMEEG/). Sources were computed with weighted minimum-norm method and dSPM (depth-weighting factor of 0.8, loosing factor of 0.2 for dipole orientation). Individual source estimate data were then projected on a template cortical surface, in order to be averaged across participants, separately

for each experiment. Mean power (i.e. square of the t-values) of regions of interest was computed to present time-courses of brain activity

1.7. Statistical analysis

1.7.1. Behavioural Data Analysis.

Behavioral data analyses were performed with Matlab software with the help of the Statistics toolbox using repeated-measures analysis for within-subject factors.

To evaluate the effect of visibility on performance and meta-performance, while factoring-out the effect of SOA more sophisticated statistical analysis was required as trial rejection and factorial analysis (SOA*Visibility*Cohorte) led to unequal number of participants in each combination of condition. Therefore, analysis of variance was performed in R software using a linear mixed-effects model ((Baayen et al., 2008) R package lme4) which allowed us to include all data available (unbalanced design) and still encompass repeated-measures. The functions used yield t statistic and, as degrees of freedom cannot be computed for this kind of analysis, p-values were derived from a Markov Chain Monte Carlo (MCMC) method.

1.7.2. MEG Data Analysis.

To detect significance differences between error and correct conditions for each type of sensor, we used a cluster-based non-parametric t-test with Monte Carlo randomization provided in the Fieldtrip software (Maris and Oostenveld, 2007). This method identifies clusters of nearby sensors presenting a significant difference between two conditions for a sufficient duration while correcting for multiple comparisons. For each sample, t-values and associated p-value were first computed by means of a Student t-test. Clusters were then

identified by taking all samples adjacent in space or in time (minimum of 2 sensors per cluster, 4.3 average spatial neighbours per EEG electrode and 8.2 per MEG channel) with $p < 0.05$. The final significance of the cluster was found by computing the sum of t-values of the entire cluster, and comparing with the results of Monte-carlo permutations (1500 permutation). Clusters were considered significant at corrected $p<0.05$ if the probability computed with the Monte-Carlo method was inferior to 2.5% (two-tailed test). As the ERN is usually observed in a 100 ms time-window after button press (Dehaene et al., 1994), cluster search was first performed on this period. To reveal more subtle differences in patients, the time-window was reduced to 30-80 ms (see Results).

For statistical analysis on a-priori clusters, average voltage over central electrodes (FC1, FC2, C1, Cz, C2) were computed over the same time-window as for the cluster analysis. Analysis was performed in Matlab using repeated-measures t-tests (two-tailed) and ANOVA with visibility and performance as within-subjects factors and group (patient versus controls) as a between-subject factor.

2. RESULTS

2.1. Schizophrenic patients are less sensitive to detect stimulus presence

We first investigated how subjective visibility varied with SOA for schizophrenic patients compared to controls (Figure 1). Subjective visibility, as measured by the percentage of *seen* responses, increased in a non-linear sigmoid manner with SOA ($F_{5,120} = 208.4$, $p < 10^{-4}$, Figure 2A), replicating earlier results (Del Cul et al., 2007; Charles et al., 2013b). Overall, no main effect of group was found ($F_{1,120}=0.09$, $p = 0.77$) and only a marginal interaction between SOA and group was found ($F_{5,120}=1.92$, $p = 0.09$) indicating no obvious change in visibility ratings in schizophrenic patients compared to controls. Interestingly though, the percentage of *seen* responses for the mask only condition was significantly higher for patients than for control subjects ($t_{24} = 2.91$, $p = 0.007$). While non significant, visibility seemed also higher in patients for short SOAs while longer SOAs were associated with lower visibility in patients, suggesting an overall lack of sensitivity for visibility reports.

To obtain a clearer idea of the ability of the patients to detect the presence of the target compared to control subjects, we transformed the raw visibility reports into an objective index of target detection sensitivity and bias, using classical signal detection theory. To this end, at each SOA level, visibility ratings (percent *Seen* responses) were compared against those in the mask-only condition, and converted to *detection-d'* and bias values (see Methods). This transformation revealed differences between patients and controls (Figure2B). While sensitivity to detect the target increased with SOA (main effect of SOA, $F_{4,96} = 205$, $p < 10^{-4}$), *detection-d'* was significantly lower for patients than for control subjects (main effect of group, $F_{1,96} = 5.28$, $p = 0.03$), especially for longer SOAs resulting in a near significant-

interaction between SOA and group ($F_{4,96} = 2.33$, $p = 0.06$). The bias towards responding that the target was absent decreased with SOA (main effect of SOA, $F_{4,96} = 205$, $p < 10^{-4}$) and was overall unchanged in patients compared to controls subject ($F_{1,96} = 0.99$, $p = 0.33$) except for the shorter SOAs for which subject were less biased towards saying that the target was absent, resulting in a near significant interaction between group and SOA ($F_{4,96} = 2.33$,$p = 0.06$). Overall these results show that subjects were less sensitive than controls to detect the presence of the target but also they were more biased to respond *"seen"* in the most uncertain conditions (shortest SOAs).

2.2. Schizophrenic patients have comparable performance in the number comparison task than controls

We then looked at the variations in performance and meta-performance as a function of SOA (Figure 2C). Objective performance in the number comparison task increased with SOA ($F_{4,96} = 108$, $p < 10^{-4}$), with a non-linear profile similar to subjective visibility (Figures 2C). Performance of the patients group were slightly lower than those of the controls overall ($F_{1,96} = 3.35$, $p = 0.08$) As intended, performance did not reach ceiling even for the largest SOA (SOA 100 ms, Figure 2C), neither for controls nor for patients group.

Next, we investigated meta-cognitive performance as a function of SOA. Our procedure allowed us to compare, on each trial, the subject's objective accuracy with his evaluation of his performance. Trials were classified as "*meta-correct*" if they were error trials perceived as errors, or correct trials perceived as correct. Otherwise they were labelled as "*meta-incorrect*". Meta-cognitive performance (i.e. percentage of meta-correct trials) increased with SOA ($F_{4,96}=149.72$, $p < 10^{-4}$), and was not significantly different across groups ($F_{1,96} = 1.62$, $p = 0.21$). Overall, these results indicate that schizophrenic patients were able to perform with normal accuracy.

2.3. Metacognitive performance of schizophrenic patients is impaired in conscious trials

We then turned to the question of how visibility affected schizophrenic patients' results in the number-comparison and the performance evaluation task. To better characterize how behaviour changed on conscious and non-conscious trials, the data were then split by visibility (*Seen* vs *Unseen*). In order to have enough trials in each category, we kept only trials corresponding to SOA larger than 33 ms for *seen* trials and those corresponding to SOA smaller than 50 ms for *unseen* trials (Charles et al., 2013b). As can be seen in figure 3A-B, both controls and schizophrenic patients performed above chance both in the number comparison task and in the performance evaluation task when they could see the target number, independently of the SOA condition (for experiments and all SOA, performance and meta-performance > 50%, $p < 0.001$).

To obtain a clearer view of relative sensitivity in the second-order performance evaluation task compared to the primary task, performance was converted to *d'* (Figures 3C) and *meta-d'* values (Figures 3D). As described by second-order Signal Detection Theory (Galvin et al., 2003; Rounis et al., 2010; Maniscalco and Lau, 2012; Charles et al., 2013b) (SDT), *d'* and *meta-d'* give an unbiased estimate of performance, respectively for first-order task (here, number comparison) and second-order task (error detection). Since these two measures are on the same scale, they allow us to compare what the first-order performance actually was to what it should have been, given second-order error detection accuracy (Galvin et al., 2003; Rounis et al., 2010; Maniscalco and Lau, 2012).

Our goal was to determine how well schizophrenic patients could perform the task and extract metacognitive information on their performance in *seen* versus *unseen* trials. Interestingly, in *seen* trials, schizophrenic patients showed decreased performance in the task compared to control subjects. While performance (*d'*, Figure 3B) and meta-performance (*meta-d',* Figure 3C) increased significantly with SOA (*d',* $F_{3,72}=100$, $p < 10^{-4}$; *meta-d',* $F_{3,72}$ =

41.2, $p < 10^{-4}$), performance was significantly lower for patients than for controls ($F_{1,72}$ = 5.79, p = 0.024). Furthermore, meta-performance seemed overall slightly lower for patients than for controls ($F_{1,96}$ = 3.84, p = 0.06), especially when considering the longest SOAs of 66 and 100 ms (t-test, all $p < 0.05$) suggesting that subjects were not able to judge their performance as well as controls for high visibility conditions. *Meta-d'* always significantly exceeded *d'* both for controls and patients ($F_{1,24}$ = 36.6, $p < 10^{-4}$) indicating that errors could be detected prior to second-order judgment, resulting in later correction of the primary judgement. However, the difference between *d'* and *meta-d'* was identical for patients and for controls ($F_{1,24}$ = 0.15, p = 0.69), resulting in an overall lower rate of detected errors for patients. This indicates that patients did not simply commit more initial errors than controls subjects that they could then detect and report as such but that the ability of patients to detect their errors was also impaired. Such result confirms impaired error-detection in patients compared to control subjects, even in maximum visibility conditions.

2.4. Cognitive and metacognitive performance are preserved on unseen trials

We next performed similar analyses of cognitive and metacognitive performance restricted to the *unseen* trials. Interestingly, in *unseen* trials schizophrenic patients performed similarly to controls. Objective performance was not significantly different in the patients and the controls group, ($F_{1,24}$=1.16, p=0.29) and increased with SOA ($F_{2,48}$=16.2, $p < 10^{-4}$). Importantly, as found before (Charles et al., 2013b), they differed from chance for SOA 50 ms (Patients,t_{12}=6,76, $p < 10^{-4}$; Controls,t_{12}=9,57, $p < 10^{-4}$) and marginally for SOA 33 ms (Patients, t_{12}=1,86, p=0.08; Controls, t_{12}=3,37, p=0.005), demonstrating a classical subliminal effect (Persaud et al., 2007; Pessiglione et al., 2007), i.e. a partial accumulation of evidence about the *unseen* targets.

Most importantly, second-order performance in the error detection task (i.e. meta-performance) was significantly above chance for both controls and patients for intermediate SOAs (SOA 33 and 50 ms, meta-performance > 50%, all $p < 0.05$). Similarly, meta-performance were not significantly different for controls and patients ($F_{1,24}$=0.196, $p = 0.66$) and increased with SOA ($F_{2,48} = 12$, $p < 10^{-4}$), confirming that even in subliminal conditions, once a primary response is emitted, participants can categorize it as correct or incorrect with better-than-chance performance, as previously found (Charles et al., 2013b).

To summarize, we found that in both patients and control subjects were above chance in judging their own errors, even on trials classified as *unseen*.

2.5. Interaction between visibility and group for performance and meta-performance

To confirm the dissociation between conscious and non-conscious processes in schizophrenia, we used a general linear model (see Methods) with SOA, visibility and decision type (first-order or second order) as within-subject factors and group as a between-subject factor. This analysis confirmed the existence of a significant interaction between group and visibility ($p = 0.056$) suggesting a specific impairment in first and second-order decisions in conscious trials for schizophrenic patients.

2.6. The error-related negativity is reduced in schizophrenic patients

We then turned to EEG recordings, in order to probe whether metacognitive performance was accompanied by an ERN and a Pe. We first investigated whether we could see an ERN in control subjects and in patients (Figure 4). We found a significant ERN,

manifested by more negative central voltages on error than on correct trials, both for controls (t_{12} = -2.63, p = 0.02) and for patients (t_{12} = -3.92 $p < 10^{-3}$) in the 0-100 ms time-window after response. Interestingly, no significant differences were found in later time-window of the Pe. Crucially, the amplitude of the ERN was not significantly different across groups (t_{24} = 0.60 p = 0.55).

We then split the data according to subjective visibility to explore how conscious perception influenced the amplitude of the ERN. Starting with the *seen* trials, a significant ERN was found for controls subjects (Figure 4B, t_{12}= -3.83, $p < 10^{-3}$) and a close-to-significance difference for patients (t_{12}= -1.89, p = 0.08). The ERN was only marginally significantly greater for controls than for patients (t_{24}= 1.49 p = 0.075). Importantly, no significant difference between correct and error was detectable on *unseen* trials for controls (t_{12} = -1.19, p = 0.26), confirming that the ERN was absent under subliminal conditions (Charles et al., 2013b). For patients a near significant ERN was found in *unseen* trials (t_{12}= -2.06 p = 0.062). The variation of the ERN with subjective report for controls was confirmed by a significant interaction between visibility (*seen* or *unseen*) and performance (*error* or *correct*) on central voltages in the time window of the ERN ($F_{1,36}$ = 11.9, p = 0.005). On the contrary for patients while this analysis revealed a main effect of performance ($F_{1,36}$ = 8.04, p = 0.015), no interaction between performance and visibility was found ($F_{1,36}$ = 0.52, p = 0.49). Interestingly, the same effect was found on the component following the ERN, the Pe which was absent for patients both in conscious (t_{12} = 0.234, p = 0.82) and in non-conscious trials (t_{12} = 0.005, p = 1) while it was clearly present for controls in *seen* trials (t_{12} = 2.38, p = 0.034) but not in *unseen* trials. Overall, these results show that the mechanisms underlying the conscious triggering of the ERN is strongly impaired in patients.

To identify the cerebral signatures of error processing, cluster analysis was applied to MEG and EEG data in order to identify any cluster of sensors showing a difference between

error and correct trials. For controls, cluster analysis essentially replicated the above ERN analysis and our previous results (Charles et al., 2013b): on *seen* trials, the typical ERN cluster on fronto-central electrodes in EEG was found ($p < 10^{-4}$, Figure 6B) while in *unseen* trials, no significant EEG cluster was detected. As found previously, some MEG sensors (magnetometers [MEGm], Figure6D) still detected a difference in activity between correct and error trials in *unseen* trials ($p = 0.023$), suggesting that distinct performance monitoring processes were still present in non-conscious trials.

For patients however, patterns of activity were very different. While the ERN topography seemed present both for *seen* (Figure 6A) and for *unseen* (Figure 6C) trials, none of the clusters reached significance, even when restricting the time-window of analysis to shorter durations. Considering MEG topographies, patterns of activity were overall very different from the one observed for controls subjects. When considering the 0-100 ms time-window, none of the clusters reached significance. However, when performing the analysis on a shorter time-window around the ERN (30-80 ms), a significant cluster of magnetometers was found active for *seen* trials ($p = 0.017$).

Statistical analysis of the difference in topography between controls and patients revealed significant differences for each sensor type, both for *seen* and *unseen* trials. In particular, in *seen* trials a significant difference was found on the ERN topography ($p<10^{-4}$), confirming that the ERN was larger for controls than for patients. Interestingly, the observed differences between control subjects and patients were not limited to voltage amplitude in identical clusters but rather reflected the existence of distinct topographies related to error in patients and in controls. In particular, the magnetometer cluster observed for patients induced a significant difference with controls group ($p<10^{-4}$) as such pattern of activity was absent for control subjects.

More surprisingly, similar results were obtained for *unseen* trials for which both EEG and MEG sensors topographies significantly differed for controls and for patients. While the the pattern of fronto-central negativity in EEG did not reached significance between controls and patients, a small set of frontal sensors were found to significantly diverge in patients compared to controls for each type of MEG sensors (all $p<10^{-4}$).

Overall, these results confirm that brain activity related to error processing in patients was not only reduced but drastically modified in patients compared to control subjects, both in conscious and non-conscious trials.

2.7. Distinct areas are involved in performance monitoring in schizophrenic patients

To shed light on the cerebral generators of the observed differences at the sensor level, we applied distributed source estimation on error and correct MEEG signals.

Our results for controls replicated our previous results (Charles et al., 2013b). For *seen* trials (Figure 6A-E), differences between error and correct trials were found bilaterally in the anterior part of the Posterior Cingulate Cortex (dPCC, MNI peak at coordinates x=-11 y=-36 z=37.5) and in dorsal anterior cingulate (dACC, MNI peak at coordinates x=-9.9 y=11.4 z=33.2). For *unseen* trials however, activation in these regions was drastically reduced. Nevertheless, as found previously, small patches in dACC (Figure 6C) remained active in the *unseen* condition, compatible with the small but significant effect detected at the sensor level in MEG data.

Interestingly however a sensibly different pattern of activity was found for patients. Considering *seen* trials, activity in all of the cingulate cortex was strongly reduced (Figure 6E). Only ventral part of cingulate cortex remained active (MNI peak at coordinates x=-3.7

y=-42.5 z=6.9, Figure 6B). This result fits with the observed pattern of activity at the sensor level, suggesting that the main generators of the ERN were inhibited in patients compared to control subjects. In *unseen* trials however, while overall activity was reduced, greater activity was found in patients compared to controls. In particular, a difference between error and correct trials was found in the rostral part of cingulate cortex (MNI peak at coordinates x=7.8 y=25,5 z=22,2, Figure 6D) as well as in the most ventral part of cingulate cortex (MNI peak at coordinates x=14.9 y=-43.9 z=-0.8) while such activity was absent in control subjects.

3. DISCUSSION

In this study we explored how meta-cognitive processes of error detection are modulated by conscious perception in schizophrenic patients. In a masking paradigm, we recorded MEEG brain responses of thirteen schizophrenic patients and thirteen age-matched control subjects to evaluate the relation between first-order performance, meta-cognition, and subjective visibility. Our findings indicate a clear dissociation between conscious and non-conscious metacognitive processes for patients: (1) Schizophrenic patients presented altered responses to error in conscious trials, manifested by a decreased ERN (2) The non-conscious computation of the likelihood of having made an error was not altered in schizophrenic patients.

Impaired metacognition and performance-monitoring processes in schizophrenia

On *seen* trials where the target number was consciously perceived, both first-order and second-order performance were significantly lower in schizophrenic patients compared to control subjects. Indeed, schizophrenic patients committed more errors than control subjects and they were also able to detect a lower fraction of them. This result fits with previous studies showing impaired metacognition in schizophrenia.

Indeed deficits in metacognition have been thought to constitute a core aspect of this pathology. Studies on meta-memory (Bacon et al., 2001; Bacon and Izaute, 2009) suggest that patients have impaired feeling of knowing (FOK) and present reduced or discordant metacognitive insight into their own memory (Bacon et al., 2001). Furthermore, impairment in theory of mind, the ability to represent and evaluate thoughts in self and others, have been hypothesized to constitute a key impairment in schizophrenic patients (Corcoran et al., 1995;

Pickup and Frith, 2001; Lysaker et al., 2011; Vargas et al., 2012), explaining both negative and positive symptoms (Pickup and Frith, 2001).

Importantly, the reduced performance in error detection was associated with a reduced amplitude of the ERN compared to control subjects. This result fits with previous findings from electrophysiological studies of the brain responses to errors in schizophrenic patients. Indeed several studies found a reduced ERN in schizophrenia (Kopp and Rist, 1999; Alain, 2002; Bates et al., 2002, 2004; Mathalon, 2002; Kim et al., 2006; Morris et al., 2006; Foti et al., 2012). Interestingly, while it was reported that the negativity after an error was reduced in schizophrenic patients compared to controls (Alain, 2002; Bates et al., 2002, 2004; Mathalon, 2002), other studies found that the ERN-like activity following correct responses, the correct response negativity (CRN), was larger in patients than in healthy controls (Alain, 2002; Mathalon, 2002; Morris et al., 2006) While such a trend was also observed in our data, it did not reach significance and comparable amplitude CRN was found in patients while decreased ERN (less negative) was indeed present. Nonetheless, such result would be consistent with our previous observation that both the CRN and the ERN are affected when a stimulus fails to reach conscious access (Charles et al., 2013b), coherent with a specific conscious deficit in schizophrenia (Del Cul et al., 2006)

The finding of a reduced ERN was coherent with the MEG results which revealed distinct patterns of activity in conscious trials for patients than for controls. Our results on control patients replicated our previous study showing that both PCC and ACC are strongly active in error compared to correct trials, with a peak of activity simultaneous to the ERN maximum. For schizophrenic patients however, only small patches of the cingulate cortex were found active, coherent with the reduced ERN amplitude. Altered performance monitoring and error responses in schizophrenia have been associated with decreased activity in prefrontal regions, particularly in cingulate cortex (Carter et al., 2001; Dehaene et al., 2003;

Laurens, 2003b; Kerns et al., 2005; Polli et al., 2008). In a similar masking paradigm, Dehaene and colleagues (Dehaene et al., 2003) showed that in conscious conditions, patients presented reduced activity in ACC in conflict trials compared to control subjects. In the present study, source reconstruction of the generators of the ERN confirms that overall activity in cingulate cortex was drastically reduced in patients compared to matched-age controls.

Altered conscious access could explain error detection deficit in schizophrenia

In a previous research, it has been proposed that the ERN is associated with the comparison or conflict between executed and intended actions (Gehring et al., 1993; Steinhauser and Yeung, 2010). In this framework, we recently proposed that error detection would result from the comparison of the output of two distinct routes: a fast non-conscious route that triggers motor action and a slow conscious route that computes intention (Charles et al., 2013a). The ERN would then reflect the conflict or the discrepancy between the outputs of these two routes. According to this view, the ERN would be tightly linked to the maintaining of goal-relevant information enabled by the triggering of the conscious route and the emergence of a conscious representation of required action.

In that respect, altered access to consciousness might indeed cause a deficit in cognitive control functions that rely on a conscious representation of information to be deployed. Indeed, deficits in action monitoring (Frith and Done, 2009; Gawęda et al., 2013) and cognitive control (Barch and Dowd, 2010; Eisenberg and Berman, 2010; Smith et al., 2011) have been largely reported in schizophrenia. Some authors proposed that these deficits were associated with specific alteration in "proactive control" (Barch and Ceaser, 2011; Lesh et al., 2013), corresponding to the maintaining of goal-relevant in preparation of a task. While according to the taxonomy proposed by Braver *et al.* (Braver, 2012), error-detection processes

should be associated preferably with "reactive control" and therefore should remain unimpaired, results of a reduced ERN in schizophrenic patients seem to contradict this theory. The hypothesis of a deficit in conscious access however might help to reconcile these views with the current findings, explaining the different deficits in cognitive control observed in schizophrenia by their tight link to conscious processing, as we showed for the ERN (Charles et al., 2013b).

Several studies found that conscious processes seem to be specifically impaired in schizophrenia while non-conscious processes are relatively preserved (Danion et al., 2001; Dehaene et al., 2003; Del Cul et al., 2006). In a recent study, Del Cul et al. showed that the priming effect manifesting, by faster reaction-time for congruent than for incongruent primes is preserved in patients (Del Cul et al., 2006), suggesting that fast automatic perceptual processes might remain intact in schizophrenic patients. However, they found that the threshold for conscious access was higher for patients, therefore speaking in favor of an alteration in cognitive processes leading to conscious access. This hypothesis was corroborated by the finding that performance in *unseen* trials in the number comparison task on the masked-prime was at chance, contrarily to control subjects in which it slightly exceeded chance. This finding suggests that schizophrenic patient were less able to exploit subliminal information below conscious threshold providing a possible explanation of their elevated threshold for conscious access.

Disctinct metacognitive processes in conscious and non-conscious conditions

In the present study, we found that schizophrenic patients were able to perform the number-comparison task slightly above chance in non-conscious conditions, with similar accuracy than controls. More importantly, they were also able to evaluate the accuracy of their own decision in unseen trials as well as control subjects. This finding, which replicate

our previous results (Charles et al., 2013b) and extend it to schizophrenic patients suggest that non-conscious processes of evaluation of the confidence in one's response can operate non-consciously. More importantly, it suggests that these processes are preserved in schizophrenia.

Crucially, this finding confirm our initial hypothesis that these non-conscious statistical evaluation processes are distinct from the one reflected by the ERN. We found previously that different patterns of activity were evoked by the ERN and by non-conscious metacognitive processes, suggesting an anatomical distinction between the two (Charles et al., 2013b). The present results confirm this hypothesis, showing that conscious mechanisms of error-detection reflected in the ERN can be impaired while non-conscious performance monitoring processes are preserved. Indeed, our results are compatible with our initial view that while the ERN reflects an all-or-none process of error detection, statistical assessment of confidence in the response can be assessed in non-conscious conditions, relying on a computationally distinct mechanism (Charles et al., 2013b).

What could be the specific alteration in conscious processing for error detection? One interesting possibility could be that schizophrenic patients present a slower access to consciousness, as a result of a higher consciousness threshold or a noisier evidence-accumulation process. According to our proposed dual-route comparison model for error-detection, slower conscious access would correspond to a slower computation of the correct response and therefore result in the unavailability of information regarding the required action even after the motor response. Interestingly, such a model predicts that the ERN would then be impeded, as no information could be used to evaluate the accuracy of the motor decision at the time of the response. In the future, insight on the dynamics of decision making and evidence accumulation, as well as analysis of the pattern of connectivity in relevant brain regions should help to test this hypothesis.

We also found that the pattern of activity in *unseen* trials associated with above-chance metacognitive performance was different from those of control subjects. In particular, activity was found in the rostral part of the cingulate cortex (rACC) as well as the frontopolar prefrontal cortex BA10 and para-hippocampic regions for schizophrenic patients. What might be the cause of these differences? One possible explanation may be that patients show increased level of engagement in the task than control subjects and therefore activate by default a different set of regions, possibly modulated differentially in error and correct trials. In particular, activity in para-hippocampic areas could reflect modification in the default level of activity in hippocampus (Heckers, 2001; Harrison, 2004) that would not be directly linked to performance monitoring. A more interesting hypothesis however could reside in the fact that as patients present impaired conscious process, accompanied with a higher threshold for consciousness, some trials associated with a high level of evidence, which would normally end up in the category of *seen* trials in controls, remain below conscious threshold in patients. Therefore, these trials could reflect a stronger level of activity for "meta-correct" trials in patients than in controls. In this respect, activity in rACC and BA10 would be expected given the previous evidence for a role of these areas in performance monitoring and confidence judgment (Fleming et al., 2010). Furthermore, such result could explain the overall increase in brain activity in *unseen* trials.

In any case, our results confirm our previous finding that a small amount of stimulus evoked activity can reach the prefrontal areas even on *unseen* trials (Charles et al., 2013b). In that respect, our findings are reminiscent of previous results showing non-conscious triggering of high-level cognitive functions linked to the monitoring of behavior such as motivation (Pessiglione et al., 2007, 2008; Capa et al., 2011), task switching (Lau and Passingham, 2007) or inhibitory processes (van Gaal et al., 2008, 2009, 2010). In accordance with a previous finding of above-chance metacognition in masking (Kanai et al., 2010), our

finding suggests that metacognitive processes linked performance monitoring can be triggered non-consciously.

Conclusion

As previously proposed, our study suggests that schizophrenia is associated with a deficit in conscious access, while non-conscious mechanisms remain largely preserved. Importantly, our findings confirm the co-existence of two different mechanisms for error detection in conscious and non-conscious conditions that are computationally distinct. While the statistical assessment of error likelihood can be deployed non-consciously and is preserved in schizophrenia, the ERN, an all-or-none conscious signal of the occurrence of an error is altered in schizophrenic patients. These findings provide new evidence on the global architecture of cognitive control and suggest new insights on the link between conscious processing and schizophrenia.

Acknowlegments

We are grateful to Narjes Bendjemaa for her help in recruiting and assessing patients; the NeuroSpin infrastructure groups, in particular to the doctors Ghislaine Dehaene-Lambertz, Caroline Huron, Lucie Hertz-Pannier and the nurses Véronique Joly-Testault, Gaëlle Mediouni Cloarec and Laurence Laurier, for their support in participant recruitment and testing; Virginie van Wassenhove, Marco Buiatti, Leila Rogeau, and all the MEG team for their help on technical difficulties;.

This project was supported by a PhD grant from the Direction Générale de l'Armement (DGA, Didier Bazalgette) and the Fondation pour la Recherche Médicale (FRM), a grant from the association Schizo-Oui and a senior grant of the European Research Council to S.D. (NeuroConsc program), as part of a general research program on functional neuroimaging of the human brain (Denis Le Bihan). The NeuroSpin MEG facility was sponsored by grants from INSERM, CEA, the Fondation pour la Recherche Médicale, the Bettencourt-Schueller foundation, and the Région île-de-France. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Alain C (2002) Neurophysiological Evidence of Error-monitoring Deficits in Patients with Schizophrenia. Cerebral Cortex 12:840–846.

Baayen RH, Davidson DJ, Bates DM (2008) Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language 59:390–412.

Bacon E, Danion JM, Kauffmann-Muller F, Bruant a (2001) Consciousness in schizophrenia: a metacognitive approach to semantic memory. Consciousness and cognition 10:473–484.

Bacon E, Izaute M (2009) Metacognition in schizophrenia: processes underlying patients' reflections on their own episodic memory. Biological psychiatry 66:1031–1037.

Barch DM, Carter CS, Braver TS, Sabb FW, MacDonald a, Noll DC, Cohen JD (2001) Selective deficits in prefrontal cortex function in medication-naive patients with schizophrenia. Archives of general psychiatry 58:280–288.

Barch DM, Ceaser A (2011) Cognition in schizophrenia: core psychological and neural mechanisms. Trends in Cognitive Sciences 16:27–34.

Barch DM, Dowd EC (2010) Goal representations and motivational drive in schizophrenia: the role of prefrontal-striatal interactions. Schizophrenia bulletin 36:919–934.

Bates AT, Liddle PF, Kiehl K a, Ngan ETC (2004) State dependent changes in error monitoring in schizophrenia. Journal of psychiatric research 38:347–356.

Bates ATA, Kiehl KK a, Laurens KR, Liddle PFP (2002) Error-related negativity and correct response negativity in schizophrenia. Clinical Neurophysiology 113:1454–1463.

Braver TS (2012) The variable nature of cognitive control: a dual mechanisms framework. Trends in Cognitive Sciences 16:105–112.

Capa RL, Bustin GM, Cleeremans A, Hansenne M (2011) Conscious and unconscious reward cues can affect a critical component of executive control. Experimental psychology 58:370–375.

Carter CS, MacDonald a W, Ross LL, Stenger V a (2001) Anterior cingulate cortex activity and impaired self-monitoring of performance in patients with schizophrenia: an event-related fMRI study. The American journal of psychiatry 158:1423–1428.

Charles L, King J-R, Dehaene S (2013a) Decoding the dynamics of action, intention, and error-detection for conscious and subliminal stimuli. In revision.

Charles L, van Opstal F, Marti S, Dehaene S (2013b) Distinct brain mechanisms for conscious versus subliminal error detection. NeuroImage 73:80–94.

Corcoran R, Mercer G, Frith CD (1995) Schizophrenia, symptomatology and social inference: investigating “theory of mind” in people with schizophrenia. Schizophrenia research 17:5–13.

Danion JM, Meulemans T, Kauffmann-Muller F, Vermaat H (2001) Intact implicit learning in schizophrenia. The American journal of psychiatry 158:944–948.

Dehaene S, Artiges E, Naccache L, Martelli C, Viard A, Schurhoff F, Recasens C, Martinot ML, Leboyer M, Sch\"urhoff F (2003) Conscious and subliminal conflicts in normal subjects and patients with schizophrenia: the role of the anterior cingulate. Proceedings of the National Academy of Sciences of the United States of America 100:13722.

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

Dehaene S, Posner MI, Tucker DM (1994) Localization of a neural system for error detection and compensation. Psychol Sci 5:303–305.

Del Cul A, Baillet S, Dehaene S (2007) Brain dynamics underlying the nonlinear threshold for access to consciousness. PLoS Biol 5:2408–2423.

Del Cul A, Dehaene S, Leboyer M, Cul A Del (2006) Preserved subliminal processing and impaired conscious access in schizophrenia. Archives of general psychiatry 63:1313.

Eisenberg DP, Berman KF (2010) Executive function, neural circuitry, and genetic mechanisms in schizophrenia. Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology 35:258–277.

Fischl B, Salat DH, van der Kouwe AJW, Makris N, Ségonne F, Quinn BT, Dale AM (2004) Sequence-independent segmentation of magnetic resonance images. NeuroImage 23 Suppl 1:S69–84.

Fleming SM, Weil RS, Nagy Z, Dolan RJ, Rees G (2010) Relating introspective accuracy to individual differences in brain structure. Science 329:1541–1543.

Fletcher P, McKenna PJ, Friston KJ, Frith CD, Dolan RJ (1999) Abnormal cingulate modulation of fronto-temporal connectivity in schizophrenia. NeuroImage 9:337–342.

Foti D, Kotov R, Bromet E, Hajcak G (2012) Beyond the Broken Error-Related Negativity: Functional and Diagnostic Correlates of Error Processing in Psychosis. Biological psychiatry 71:864–872.

Friston K (2005) Disconnection and cognitive dysmetria in schizophrenia. The American journal of psychiatry 162:429–432.

Friston K, Frith CD (1995) Schizophrenia: a disconnection syndrome. Clin Neurosci.

Friston KJ (1998) The disconnection hypothesis. Schizophrenia research 30:115–125.

Frith CD, Done DJ (2009) Experiences of alien control in schizophrenia reflect a disorder in the central monitoring of action. Psychological Medicine 19:359.

Galvin SJ, Podd J V, Drga V, Whitmore J (2003) Type 2 tasks in the theory of signal detectability: discrimination between correct and incorrect decisions. Psychonomic bulletin & review 10:843–876.

Gawęda L, Woodward TS, Moritz S, Kokoszka A (2013) Impaired action self-monitoring in schizophrenia patients with auditory hallucinations. Schizophrenia research.

Gehring WJ, Goss B, Coles MGH, Meyer DE, Donchin E (1993) A neural system for error detection and compensation. Psychological Science 4:385–390.

Grillon M-L, Oppenheim C, Varoquaux G, Charbonneau F, Devauchelle A-D, Krebs M-O, Baylé F, Thirion B, Huron C (2012) Hyperfrontality and hypoconnectivity during refreshing in schizophrenia. Psychiatry research 211:226–233.

Haraldsson H (2004) Transcranial Magnetic Stimulation in the investigation and treatment of schizophrenia: a review. Schizophrenia Research 71:1–16.

Harrison PJ (2004) The hippocampus in schizophrenia: a review of the neuropathological evidence and its pathophysiological implications. Psychopharmacology 174:151–162.

Heckers S (2001) Neuroimaging studies of the hippocampus in schizophrenia. Hippocampus 11:520–528.

Huddy VC, Aron a R, Harrison M, Barnes TRE, Robbins TW, Joyce EM (2009) Impaired conscious and preserved unconscious inhibitory processing in recent onset schizophrenia. Psychological medicine 39:907–916.

Hughes ME, Fulham WR, Johnston PJ, Michie PT (2011) Stop-signal response inhibition in schizophrenia: Behavioural, event-related potential and functional neuroimaging data. Biological psychology 89:220–231.

Kanai R, Walsh V, Tseng C-H (2010) Subjective discriminability of invisibility: A framework for distinguishing perceptual and attentional failures of awareness. Consciousness and cognition 19:1045–1057.

Kerns JG, Cohen JD, Macdonald AM, Johnson MK, Stenger VA, Aizenstein HJ, Carter CS (2005) Decreased conflict- and error-related activity in the anterior cingulate cortex in subjects with schizophrenia. The American journal of psychiatry 162:1833–1839.

Kim M-S, Kang SS, Shin KS, Yoo SY, Kim YY, Kwon JS (2006) Neuropsychological correlates of error negativity and positivity in schizophrenia patients. Psychiatry and clinical neurosciences 60:303–311.

Kopp B, Rist F (1999) An event-related brain potential substrate of disturbed response monitoring in paranoid schizophrenic patients. Journal of abnormal psychology 108:337–346.

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

Laurens KR (2003a) Rostral anterior cingulate cortex dysfunction during error processing in schizophrenia. Brain 126:610–622.

Laurens KR (2003b) Rostral anterior cingulate cortex dysfunction during error processing in schizophrenia. Brain 126:610–622.

Lesh T a., Westphal AJ, Niendam T a., Yoon JH, Minzenberg MJ, Ragland JD, Solomon M, Carter CS (2013) Proactive and reactive cognitive control and dorsolateral prefrontal cortex dysfunction in first episode schizophrenia. NeuroImage: Clinical.

Liang M, Zhou Y, Jiang T, Liu Z, Tian L, Liu H, Hao Y (2006) Widespread functional disconnectivity in schizophrenia with resting-state functional magnetic resonance imaging. Neuroreport 17:209–213.

Lysaker PH, Olesek KL, Warman DM, Martin JM, Salzman AK, Nicolò G, Salvatore G, Dimaggio G (2011) Metacognition in schizophrenia: correlates and stability of deficits in theory of mind and self-reflectivity. Psychiatry research 190:18–22.

Maniscalco B, Lau HC (2012) A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. Consciousness and cognition 21:422–430.

Maris E, Oostenveld R (2007) Nonparametric statistical testing of EEG- and MEG-data. Journal of neuroscience methods 164:177–190.

Mathalon DH (2002) Response-monitoring dysfunction in schizophrenia: an event-related brain potential study. Journal of abnormal ….

Morris SE, Yee CM, Nuechterlein KH (2006) Electrophysiological analysis of error monitoring in schizophrenia. Journal of abnormal psychology 115:239–250.

Persaud N, McLeod P, Cowey A (2007) Post-decision wagering objectively measures awareness. Nature Neuroscience 10:257–261.

Pessiglione M, Petrovic P, Daunizeau J, Palminteri S, Dolan RJ, Frith CD (2008) Subliminal instrumental conditioning demonstrated in the human brain. Neuron 59:561–567.

Pessiglione M, Schmidt L, Draganski B, Kalisch R, Lau HC, Dolan RJ, Frith CD (2007) How the brain translates money into force: a neuroimaging study of subliminal motivation. Science 316:904.

Pickup GJ, Frith CD (2001) Theory of mind impairments in schizophrenia: symptomatology, severity and specificity. Psychological medicine 31:207–220.

Polli FE, Barton JJS, Thakkar KN, Greve DN, Goff DC, Rauch SL, Manoach DS (2008) Reduced error-related activation in two anterior cingulate circuits is related to impaired performance in schizophrenia. Brain 131:971–986.

Rounis E, Maniscalco B, Rothwell JC, Passingham RE, Lau HC (2010) Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. Cognitive Neuroscience 1:165–175.

Schmitt A, Hasan A, Gruber O, Falkai P (2011) Schizophrenia as a disorder of disconnectivity. European archives of psychiatry and clinical neuroscience 261 Suppl :S150–4.

Smith EE, Eich TS, Cebenoyan D, Malapani C (2011) Intact and impaired cognitive-control processes in schizophrenia. Schizophrenia research 126:132–137.

Steinhauser M, Yeung N (2010) Decision processes in human performance monitoring. The Journal of Neuroscience 30:15643–15653.

Tononi G, Edelman GM (2000) Schizophrenia and the mechanisms of conscious integration. Brain research Brain research reviews 31:391–400.

Uhlhaas PJ, Singer W (2010) Abnormal neural oscillations and synchrony in schizophrenia. Nature reviews Neuroscience 11:100–113.

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, Scholte HS, Lamme V a F (2010) Unconscious activation of the prefrontal no-go network. The Journal of Neuroscience 30:4143.

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.

Vargas M, Sendra J, Benavides C (2012) Metacognitive Dysfunction in Schizophrenia. Edited by THJ Burne.

Yan H, Tian L, Yan J, Sun W, Liu Q, Zhang Y-B, Li X-M, Zang Y-F, Zhang D (2012) Functional and anatomical connectivity abnormalities in cognitive division of anterior cingulate cortex in schizophrenia. PloS one 7:e45659.

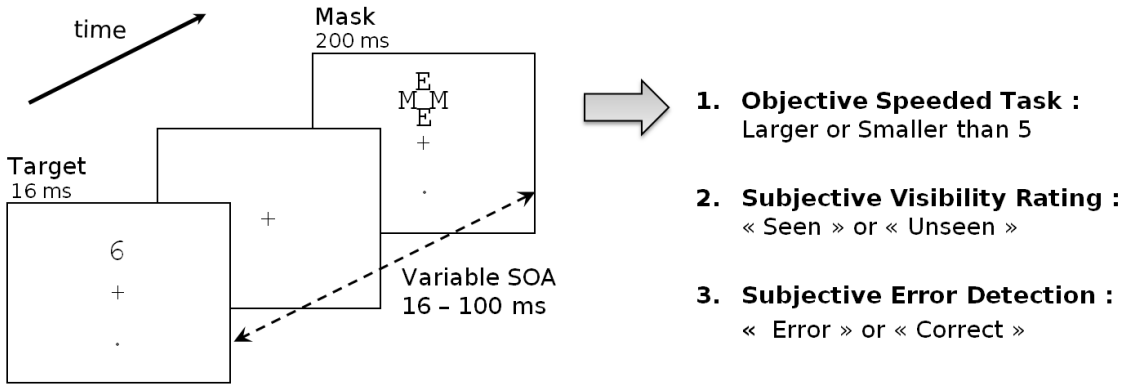


Figure 1: Experimental design.

On each trial, a number was presented for 16 ms at one of two possible locations (top or bottom). It was followed by a mask composed of a fixed array of letters centered on the target location. The delay between target onset and mask onset (SOA) varied randomly across trials (16, 33, 50, 66 or 100 ms). In one sixth of the trials, the mask was presented alone (mask only condition). Participants first performed an objective forced-choice number comparison task where they decided whether the number was smaller or larger than five. For this task, participant were instructed to respond as fast as they could while maintaining accuracy. Then, on each trial, participants performed two subjective tasks. First they evaluated the subjective visibility of the target by choosing between the words “Seen” and “Unseen”, displayed randomly either left or right of fixation. Second, they evaluated their own performance in the primary number comparison task by choosing between the words “Correct” and “Error”, again displayed randomly either left or right.

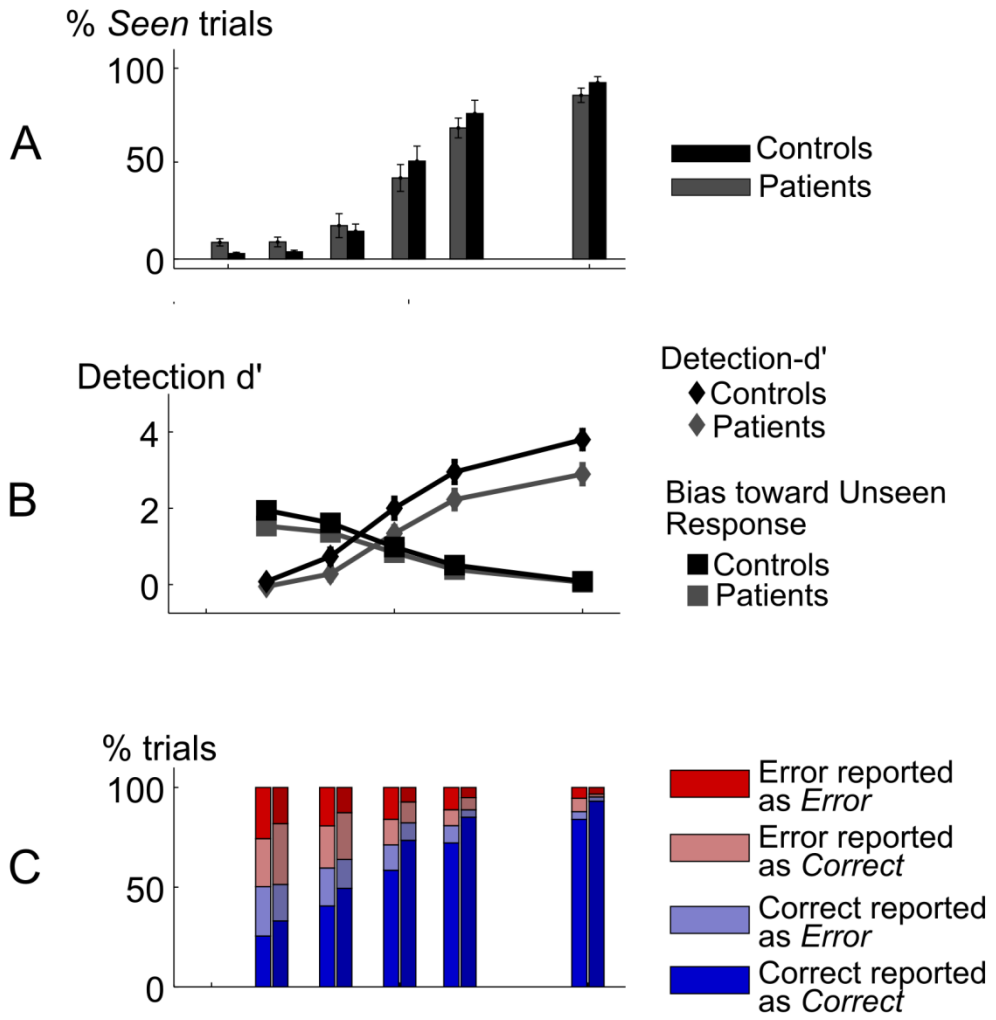


Figure 2: Visibility and performance results according to SOA for Patients and Controls.

(A) Visibility ratings, expressed as the proportion of seen responses as a function of SOA. Gray bars correspond to patient data and black bars correspond to controls data.

(B) The circle points represent detection-d′ values while the squares represents response bias towards unseen response (same scale as detection-d′), for each SOA. Gray lines correspond to patient data and black bars correspond to controls data.

(C) Percentage of each category of trials according to actual objective performance and subjective report of performance (Error trials correctly classified as Error in dark red, Correct trials correctly classified as Correct in dark blue, Error trials incorrectly classified as Correct, in light red and Correct trials incorrectly classified as Error in light blue) for each SOA. For each set of bars, left bars correspond to patients data while right bar correspond to controls data.

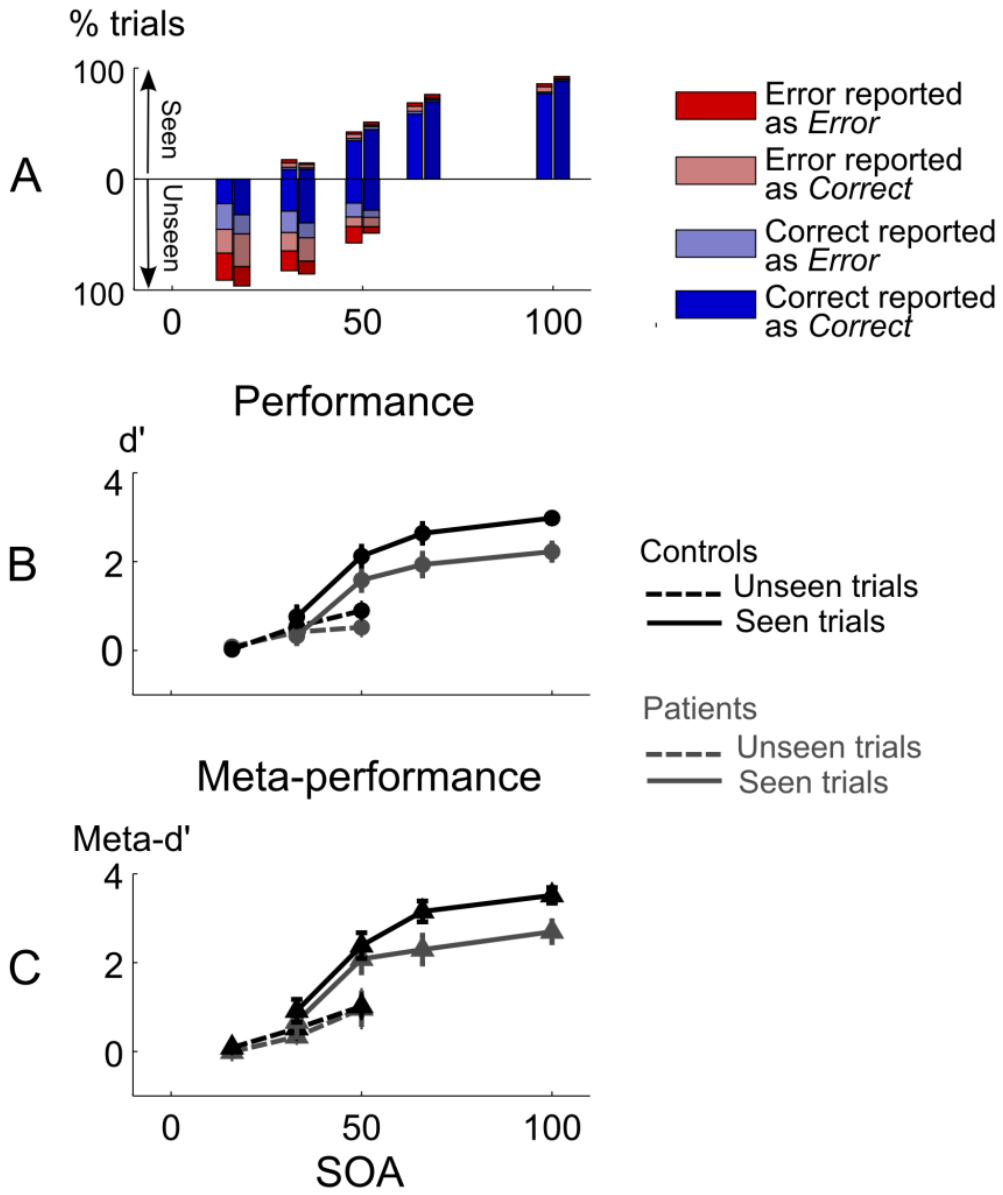


Figure 3: Performance and meta-performance according to visibility and SOA for patients and controls

(A) Proportions of unseen (below midline) and seen trials (above midline) were computed for each SOA. For each type of trials and each SOA, the relative percentage of each category of trials was derived according to objective performance and subjective report of performance (same color code as in Fig. 2).

(B) Unbiased measures of performance (d′, circles) for controls (black line) and patients (gray lines) were computed separately for seen (solid line) and unseen (dashed-line) trials and each SOA value. All error-bars represent standard error.

(C) Unbiased measures of performance (meta-d′, triangles) for controls (black line) and patients (gray lines) were computed separately for seen (solid line) and unseen (dashed-line) trials and each SOA value. All error-bars represent standard error.

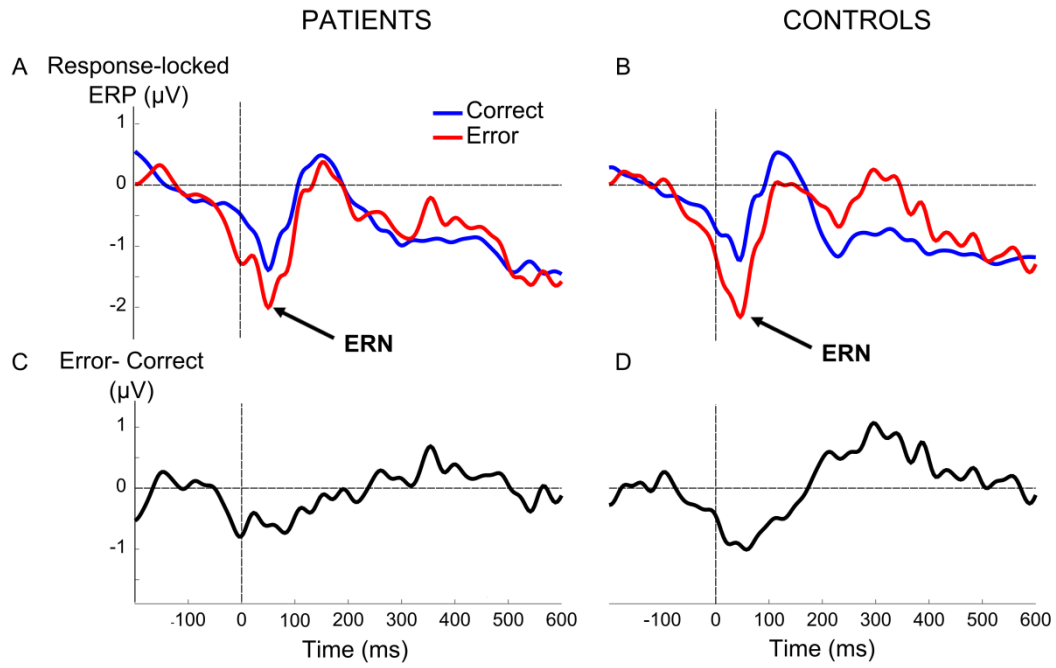


Figure 4: Time courses of event-related potentials as a function of objective performance for controls and patients.

(A-B) Grand-average event-related potentials (ERPs) recorded from a cluster of central electrodes (FC1, FC2, C1, Cz, C2), sorted as a function of whether performance was erroneous (red lines) or correct (blue lines), for patients (left panel) and controls (right panel)

(C-D) Difference waveforms of error minus correct trials for patients (left panel) and controls (right panel)

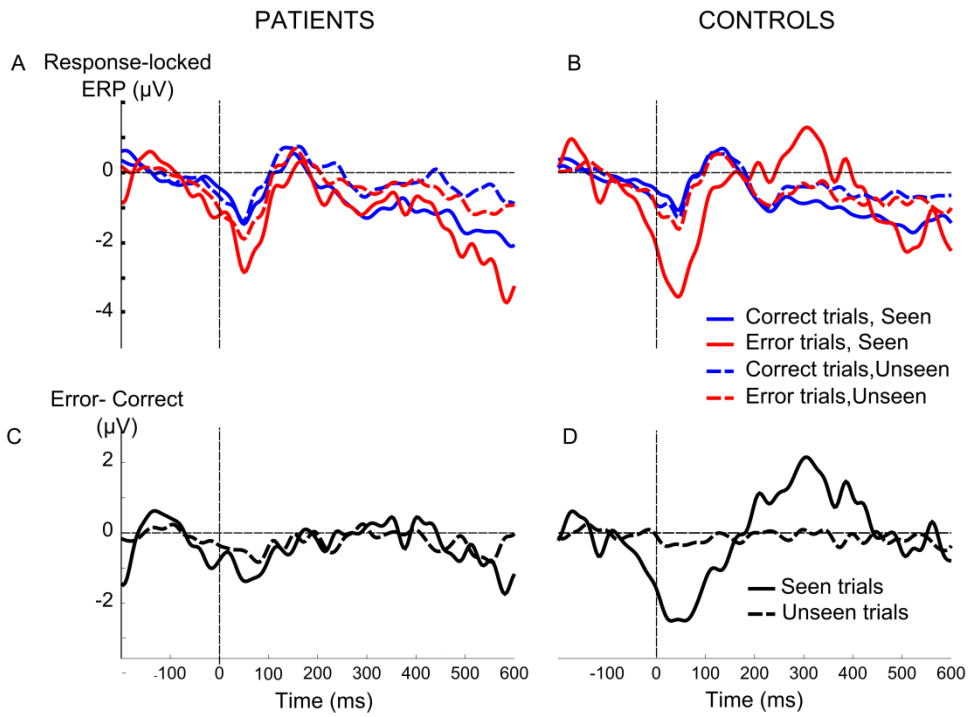


Figure 5: Time courses of event-related potentials as a function of objective performance and visibility for controls and patients.

(A-B) Grand-average event-related potentials (ERPs) recorded from a cluster of central electrodes (FC1, FC2, C1, Cz, C2), sorted as a function of whether performance was erroneous (red lines) or correct (blue lines), for patients (left panel) and controls (right panel), for *seen* (solid lines) and *unseen* (dashed lines) trials.

(C-D) Difference waveforms of error minus correct trials for patients (left panel) and controls (right panel), for *seen* (solid lines) and *unseen* (dashed lines) trials.

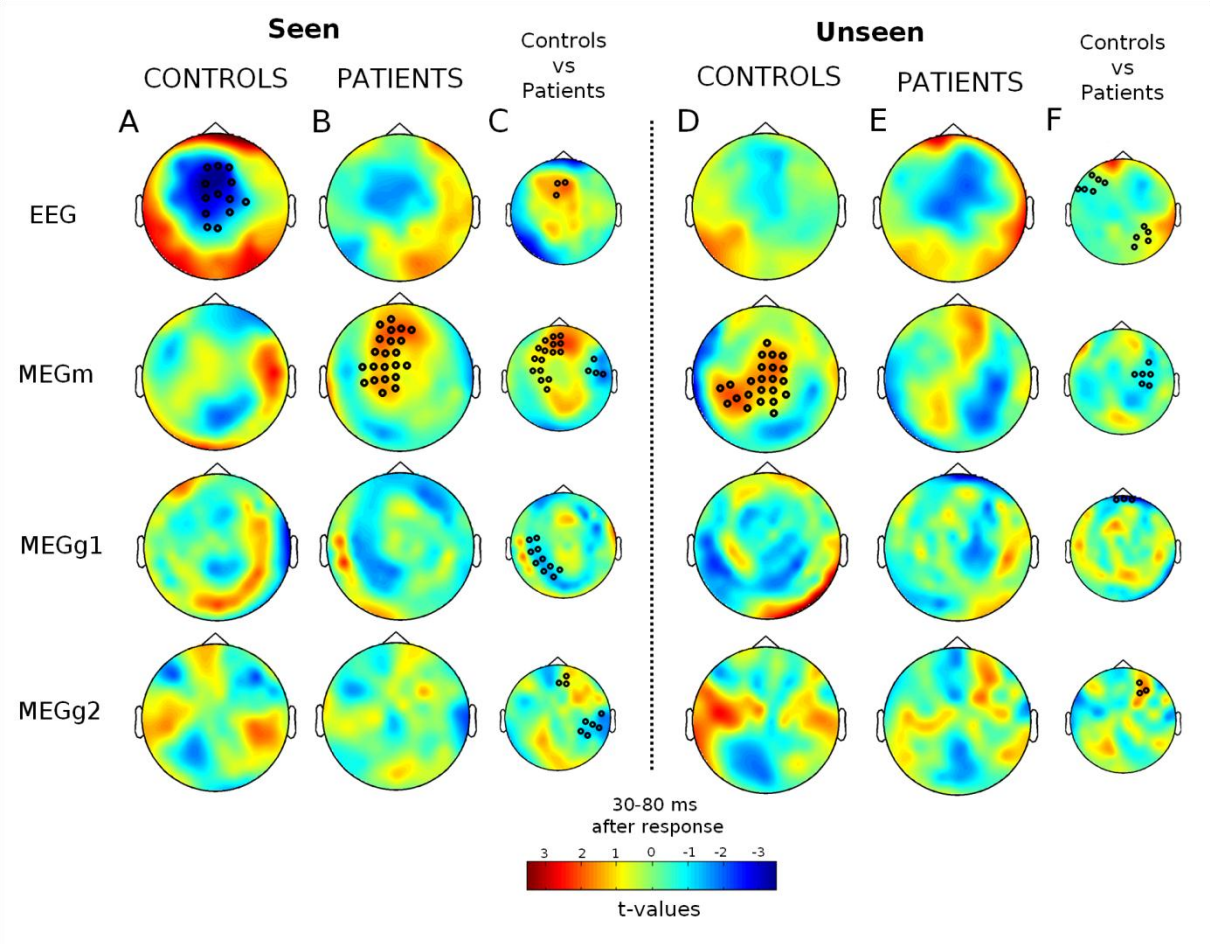


Figure 6: Error-related MEEG topographies as a function of target visibility for patients and controls.

(A-B, D-E) Each plot depicts the scalp topography of the t-value for a difference between correct and error trials, averaged across a 30–80 ms time window following the motor response, separately for each type of sensors (EEG, magnetometers, [MEGm], longitudinal gradiometers [MEGg1], latitudinal gradiometers [MEGg2]) and for *seen* and *unseen* trials, for controls (A,D) and patients (B,E). Black circles indicate sensors belonging to a spatiotemporal cluster showing a significant difference ($p < 0.025$) between error and correct conditions using a Monte-Carlo permutation test.

(C,F) Each plot depicts the scalp topography of the t-value for a difference between patients and controls trials for the subtraction error-correct, averaged across a 30–80 ms time window following the motor response, separately for each type of sensors (EEG, magnetometers, [MEGm], longitudinal gradiometers [MEGg1], latitudinal gradiometers [MEGg2]) and for *seen* (C) and *unseen* (F) trials. Black circles indicate sensors belonging to a spatiotemporal cluster showing a significant difference ($p < 0.025$) between error and correct conditions using a Monte-Carlo permutation test.

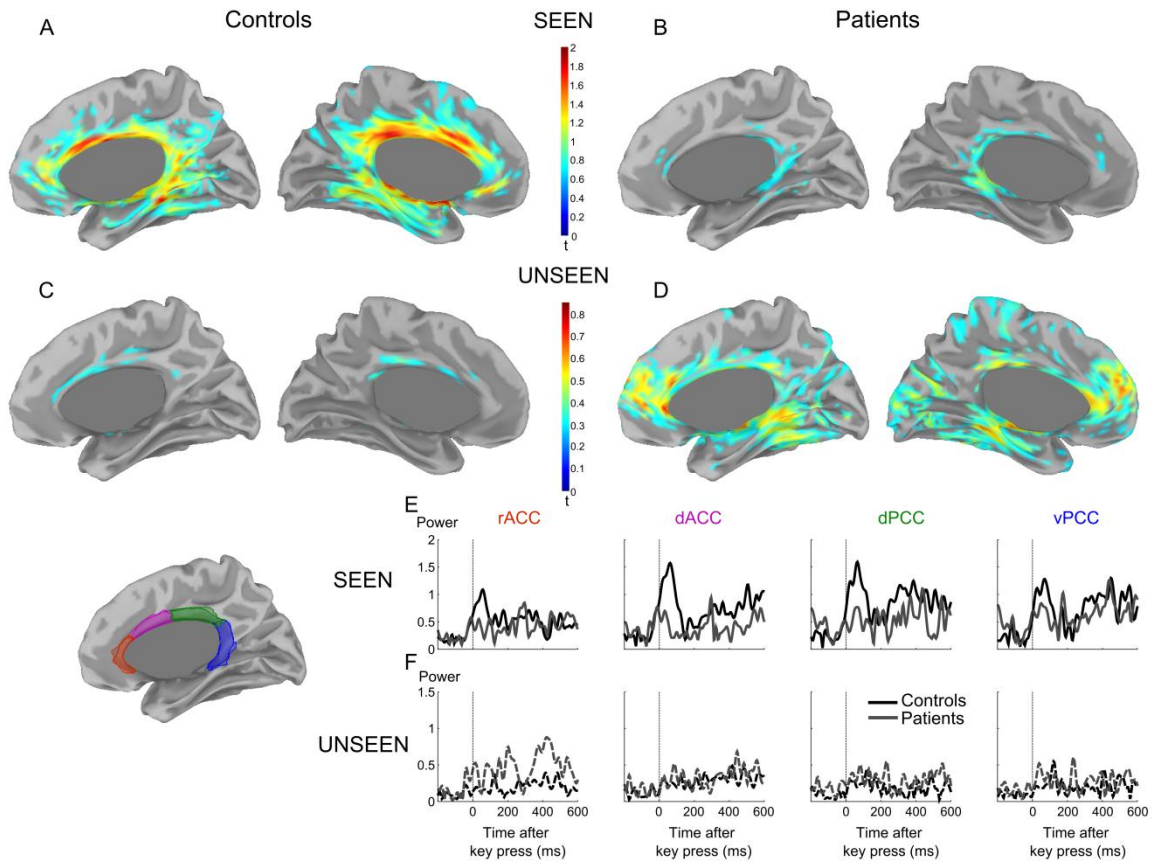


Figure 6: Difference of source estimates between error and correct trials for patients and controls.

(A–D) View of the medial surface of the left and right hemispheres, for controls (A,C) and patients (B,D), in *seen* (A–B) and *unseen* (C–D) trials. Data are thresholded at 33% of maximum activity within each condition. Brain activity was averaged in a 0–100 ms time-window.

(E–F) Time course of brain activity in four bilateral regions of interest located in ventral Anterior Cingulate Cortex (vACC), dorsal Anterior Cingulate Cortex (dACC), dorsal Posterior Cingulate Cortex (dPCC) and ventral Posterior Cingulate Cortex in *seen* (E, solid lines) and *unseen* (F, dotted lines) trials, for patients (gray lines) and controls (black lines).

Values correspond to instantaneous power in the region of interest (average, across vertices, of the square current density t-maps).

Part III

General discussion

The work contained in this thesis is concerned with analyzing the relationship between consciousness and metacognition. We have focused on a specific metacognitive task, error-detection, to assess whether metacognitive information can be extracted in non-conscious conditions. Our findings suggest that distinct metacognitive processes are triggered consciously and non-consciously. We found that only when we consciously see a target stimulus do we establish a stable representation of the associated required action that can be further decoded in patterns of brain activity. Crucially, we showed that this representation triggers evaluation processes, associated with activity in brain areas related to performance monitoring, that allow assessment of the accuracy of our own actions. We proposed a model of meta-decision in which conscious error detection results from the comparison of actual motor response with the conscious representation of the required action, a process that we found to be correlated in time and in amplitude with the amount of information available on each decision. We showed that this process however was distinct from the one triggered in non-conscious conditions. While the above-described high-level abstract representations of required action and all-or-none performance evaluation could not be found in non-conscious trials, we nevertheless showed that some action monitoring processes existed for those trials, allowing prediction above-chance level of the accuracy of decisions. Importantly, these processes were distinct from that observed in conscious conditions as we show they were preserved in patients presenting a deficit in conscious error detection. We propose a computational model of conscious and non-conscious decision to account for our findings.

Below we will discuss contributions, limitations, and future perspectives relevant to each of the questions raised in the thesis - the impact of such findings on the field of research of consciousness, what such findings suggest concerning the models of decisions and meta-decisions and the new perspective of research suggested by this work.

CHAPTER 7

Implications for the models of consciousness

7.1 The depths of non-conscious processes revisited

We have seen in the introduction of this manuscript that several executive processes can be initiated without consciousness. Implicit learning can occur non-consciously (Destrebecqz and Cleeremans, 2001) and can trigger complex motivational processes (Pessiglione et al., 2008; Pessiglione et al., 2007; Schmidt et al., 2010; Capa et al., 2011). Moreover, subliminal information about task-set can influence task performance and related brain activity (Lau and Passingham, 2007; De Pisapia et al., 2011; Reuss et al., 2011; Zhou and Davis, 2012; Mattler, 2003; Martens et al., 2011). Similarly, it was shown that inhibitory control mechanisms can be triggered by non conscious stimuli (Cohen et al., 2009; van Gaal et al., 2009), activating related regions of prefrontal cortex (van Gaal et al., 2010; van Gaal et al., 2008). Finally, some performance monitoring processes seem to be triggered non-consciously (Logan and Crump, 2010; Nieuwenhuis et al., 2001; Endrass et al., 2007; Cohen et al., 2009).

In the present work, we showed that some processes related to metacognitive judgments operate outside of consciousness. Our results concur with those of a recent study (Kanai et al., 2010) in which metacognitive abilities were tested in different masking paradigms, either manipulating the amount of information concerning the stimulus (weak contrast, backward masking or continuous flash suppression) or manipulating the attentional resources (dual-task, attentional blink or high spatial uncertainty). Subjects performed a detection task on the masked stimuli but also reported the confidence they had in their response in a binary manner. The authors showed that indeed, in such detection tasks, the II-order AUC was above chance for all masking conditions. Interestingly, only a second measure based on trials in which the target was absent revealed a significant difference between conditions in which attention was manipulated, compared to classic masking conditions. However, as these results were based simply on a detection task, assessing the presence or the absence of the masked target, it was difficult to determine if such ability could extrapolate to more complex paradigms. Moreover, the task used was not orthogonal to conscious perception, therefore making it difficult to interpret if the obtained results consisted in a form of metacognition on conscious access or whether it reflected true non-conscious response monitoring process.

In the experimental work of this thesis, we were able to disentangle these two aspects by showing that some performance-monitoring processes could be triggered when subjects denied seeing the target

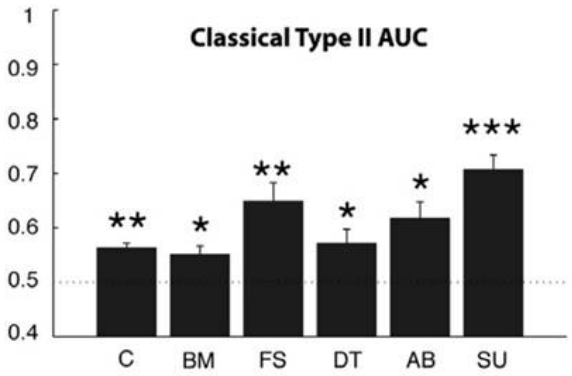


Figure 7.1: Type II AUC reflecting sensitivity in confidence judgements in different masking paradigms from Kanai et al., 2010. For six different subliminal perception paradigms (C=contrast reduction, BM=backward masking, FS=flash suppression, DT=dual task, AB=attentional blink and SU=spatial uncertainty), Type II performance was computed. Type-I performance was kept constant across all experiments. In each tasks, results reveal that type II performance was significantly greater than chance.

stimulus. It therefore suggests that metacognition judgments concerning our own motor response can be partially deployed non-consciously. This finding is reminiscent of reports in blindsight patients showing that not only can these patients perform above chance on many tasks when stimuli are presented in their blind hemi-field, but that they can also provide relatively accurate confidence judgments on their response (Evans and Azzopardi, 2007).

Interestingly, this finding was replicated in a population of schizophrenic patients. In conscious trials, patients presented impaired meta-cognitive performance associated with decreased activity in the cingulate region. In non-conscious trials however, meta-performance was identical to those of control subjects, reproducing our earlier finding of above-chance confidence judgments. This dissociation suggests that the underlying metacognitive process of this "blindsight" effect is truly distinct from the one at stake in conscious perception. Crucially, this effect was only observed for intermediate conditions of SOA for which performance was already above-chance. Indeed our initial finding showed that for the shortest conditions of SOA in which both target detection and performance were at chance, meta-performance also remained at chance level. Only for longer SOAs when performance improved, did meta-performance increase in the same manner, showing that the information used for first-order decision may be used in a second-order judgment as well.

What is the impact of such a finding regarding the depth of non-conscious processing? Six years ago, Dehaene et al. (2006) proposed a taxonomy to classify the range of non-conscious processes and characterize how these processes were modulated by attention (Figure 7.2). In particular, they proposed that while complete subliminal conditions in which the stimulus is unattended would be characterized by very weak activity in early visual cortex and almost no priming effect, subliminal stimuli that are attended would present significantly enhanced activity. However, while these stimuli would

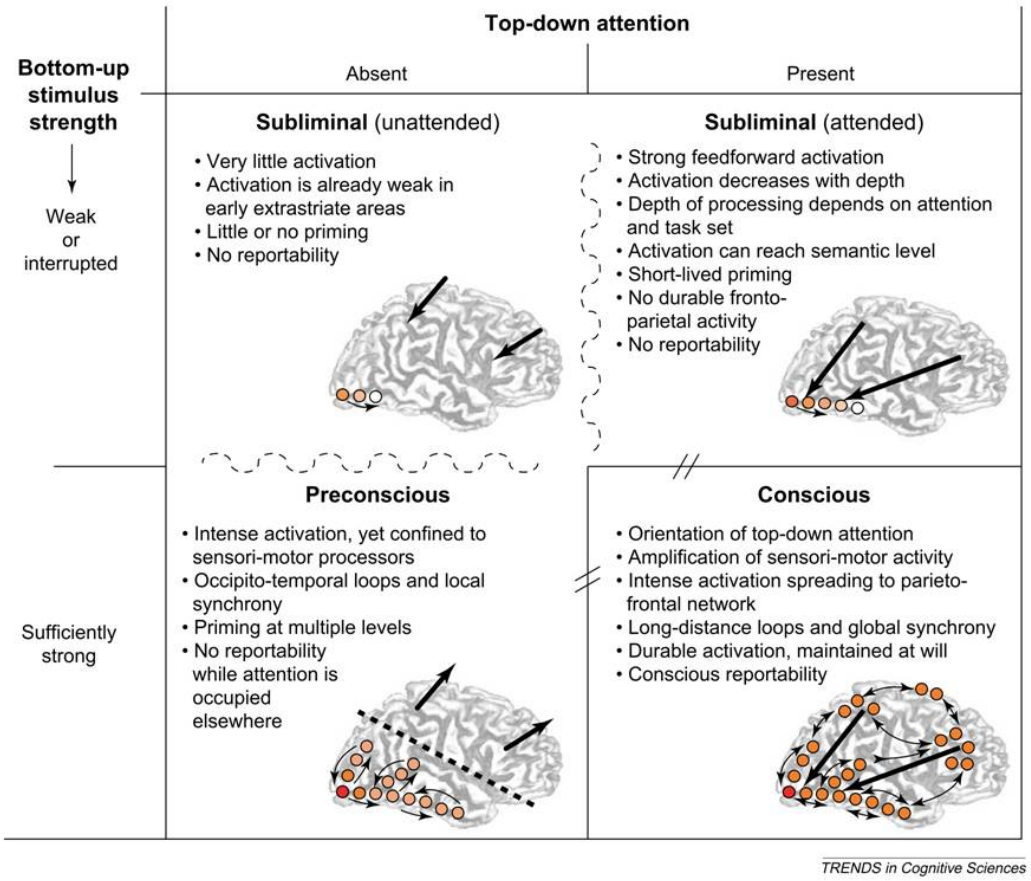


Figure 7.2: Taxonomy of subliminal, preconscious and conscious processes (from Dehaene et al., 2006)

elicit stronger feed-forward activation, evidence would nonetheless be progressively diluted with the depth of cognitive processing, preventing the crossing of the threshold for conscious access. In contrast, conscious processing would be characterized by intense activation spreading to parieto-frontal networks and could be sustained in time by long-range connections and synchrony between distant brain areas. Crucially, the authors also proposed the existence of a third category of stimuli that would be characterized by strong activation but would not be perceived consciously due to a failure in attentional processes. According to their previous results (Sergent et al., 2005; Sergent and Dehaene, 2004a; Sergent and Dehaene, 2004b), this type of stimuli could be associated with strong priming effects due to sufficient input evidence but yet would stay confined to sensori-motor processors, information remaining encapsulated and failing to trigger durable fronto-parietal activity due to the "bottleneck" for attentional resources.

At first glance, our findings appear difficult to reconcile with the above-described model. It has been shown that stimuli that are presented at sensory threshold may not be perceived depending on the state of vigilance or the ongoing activity before onset of the stimulus (Linkenkaer-Hansen et al., 2004; Busch

et al., 2009; Wyart and Sergent, 2009), while being associated nonetheless with partial accumulation of evidence. However, in the taxonomy proposed initially (Dehaene et al., 2006), this type of perception should not be associated with above-chance second-order metacognitive judgment. In particular, it has been suggested that the characteristics of subliminal processing compared to preconscious state is a short decay in time, with the effect of subliminal perception lasting only a few hundred milliseconds. Our results are at odds with this view as we show that introspection on *unseen* stimuli can last up to two seconds after stimulus presentation. In this regard, our condition of stimulation seems to be more similar to preconscious states in which stimulation would have the strength to elicit a conscious percept in optimal condition for perception, but fails to cross the threshold for conscious access. Indeed it has been suggested that in some metacontrast masking conditions, visibility can be better predicted by state of connectivity between V1 and higher visual areas than simply by the level of activation (Haynes et al., 2005) in primary visual areas. As in intermediate SOAs, approximately half of the trials are reported as *seen* and half as *unseen*, one can imagine that the level of evidence about the stimulus is close to threshold and therefore small variations in the state of brain activity can have a strong impact on further visual process. In this respect, such a mechanism could have the same effect as attentional mechanisms, both masking and inattention producing in some circumstances a state that can be qualified as "preconscious". It has been argued that indeed such a state is particularly relevant for the study of conscious processing as it allows one to overcome an important confound of consciousness research which is that non-conscious processes are often associated with much weaker input signal than conscious processes (Lau, 2012).

7.2 Crossing of the threshold for conscious access : an all-or-none phenomenon ?

Interestingly we found that although metacognitive processes may operate non-consciously, crossing the threshold for conscious perception has a major impact on brain activity and its response to errors. In particular, we found that the ERN, a well-known brain marker of performance monitoring, was evoked solely for conscious stimuli. This finding was independent of time-pressure, as an ERN was observed in two distinct experiments where the main task was either very strongly or moderately speeded. Moreover, conscious access did not influence solely the response to error but also had an impact on the negativity following correct trials, confirming the fact that crossing the threshold for conscious access induced important modifications in brain processes related to response monitoring.

This finding is in accordance with prediction of the Global Neuronal Workspace (GNW) model (Dehaene et al., 2006; Dehaene and Changeux, 2011; Sergent and Dehaene, 2004b; Baars, 2002; Dehaene and Naccache, 2001) which postulates that conscious access is associated with a sharp transition in brain activity allowing for information content to be broadly broadcasted and to form a sustained conscious representation. Importantly, this process would rely on long-range connectivity between distant brain areas, in particular prefrontal, cingulate, and parietal regions, allowing for specialized non-

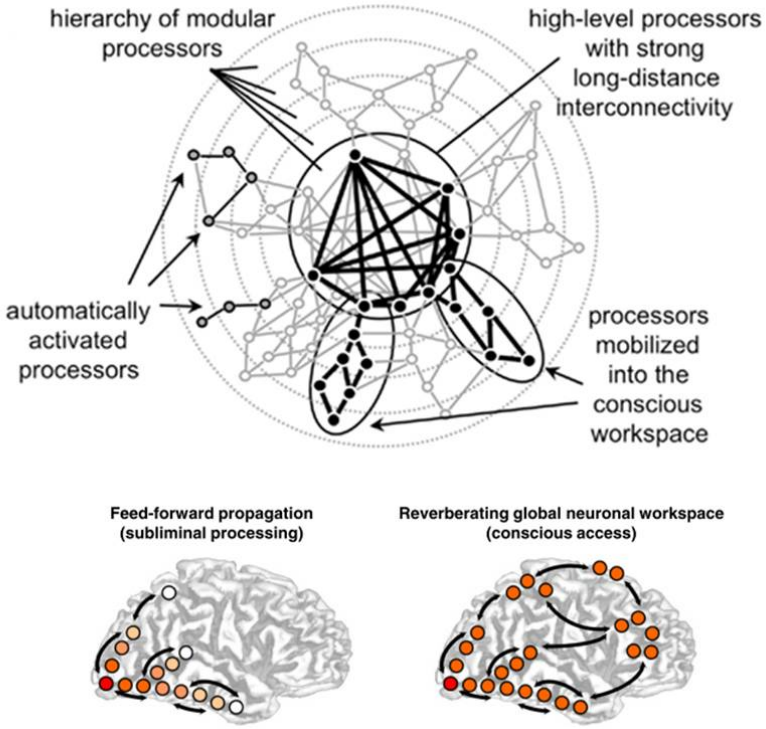


Figure 7.3: The Global Neuronal Workspace model (from Dehaene et al., 2006 and Dehaene and Changeux, 2011)

conscious processors to inter-connect through this global workspace (Figure 7.3). The pattern of activity encoded by connected assembly would allow the maintenance of a specific neural representation that could then be further processed by other specialized processors and be verbally reported. According to this model, although non-conscious stimuli could be partly processed along automatized processing routes, they would remain encapsulated to a set of brain regions; limiting their processing to further stages of processing.

One crucial prediction of the GNW is that we should find markers of the ignition process corresponding to access to a sustained and global availability of information. Crucially, these markers should behave in the same non-linear way as verbal reports, their emergence being tightly linked to subjective conscious experience. Indeed, our findings demonstrate the existence of such markers as we found that the ERN indexing response monitoring process encompasses these predictions. Indeed, we found that the ERN varied in an all-or-none fashion with subjective visibility, beyond the variations in stimulus strength, performance and meta-performance.

Importantly, while the ERN varied in the same manner as subjective report, it is difficult to imagine that it plays a role in conscious access. In that sense, the ERN does not reflect a true neural correlate of consciousness (**NCC**) but rather, one of its consequences, highlighting how conscious access enables the triggering of further processes, in particular those linked to behaviour monitoring. As proposed by

the taxonomy developed by Aru et al. (2012) which separates true NCC from its prerequisites and its consequences, the ERN might fall into the latter category, constituting as an NCC-co a by-product of a true NCC.

Also worthy of note is the fact that the ERN does not constitute a true correlate of conscious error detection. Indeed, the presence of an ERN does not necessarily mean that the error will be consciously perceived (Nieuwenhuis et al., 2001; Endrass et al., 2007). The Pe, the positive component following the ERN might in this respect be a more plausible candidate as a true neural correlate of error awareness (Nieuwenhuis et al., 2001; Steinhauser and Yeung, 2010). Therefore while the ERN might constitute an NCC-co for perceptual awareness, it might constitute a prerequisite (NCC-pr) for error awareness. This view completes the original GNW model by showing that conscious processes might sometimes be embedded in one another, constructing a global architecture of processing linked to conscious access of different objects.

Crucially, we were able to show that the presence of the ERN was determined by the emergence of a representation of the correct response (decodable in brain activity) that served as a comparison point for the actual response in order to evaluate the accuracy of the motor decision. This representation was present only in the conscious condition, when the subject reported consciously perceiving the target. Importantly, it was independent of the motor response made by the subject and its accuracy, suggesting that it reflected a high-level representation that is related to the conscious perception of the required action and distinct from the ongoing action plan. We argued that this representation constitutes a conscious intention signal (Desmurget and Sirigu, 2012; Desmurget et al., 2009) that might sometimes arrive too late to directly modulate action but might play a key-role in the evaluation and monitoring of actions. Interestingly, we found that this representation varied also in an all-or-none fashion with subjective reports of visibility, being one of the possible substrate of conscious decision on the stimulus.

Are these findings compatible with an all-or-none view of consciousness? We showed that indeed some processes seem to be indexed to conscious visibility report, following their all-or-none variation. However, we saw that some metacognitive computations can occur outside of consciousness. Furthermore, our findings suggest that even for consciously accessed stimuli, evidence can continue to be accumulated after the crossing of the threshold for conscious access, as reflected by our result on the variation of the ERN with SOA. Indeed, we found that in seen trials, error are detected better according to masking strength and the level of information on the correct responses. Therefore, our findings suggest that while some processes indeed reflect an all-or-none aspect of conscious perception, other processes vary more continuously with the objective level of information that enters the system. Indeed, while conscious access might constitute in itself a discrete process, corresponding to a non-linear step in the global sharing and the availability of information, conscious content on the other hand might still vary in its level of evidence, reflecting a continuous state of accumulation of evidence. Nonetheless, consciousness might often be accompanied by a large improvement in the evidence accumulation process, explaining the sudden increase in performance in conscious conditions.

7.3 Implication for the measure of consciousness

What impact do our findings have on the question of how to measure consciousness? Our results show that above-chance performance can be obtained in forced-choice metacognitive tasks for unconscious trials. Therefore, they seem incompatible with the use of wagering or confidence judgment as an index of conscious perception (Kunimoto et al., 2001; Persaud et al., 2007).

The wagering technique has been proposed to measure awareness in an optimal way, truly reflecting the conscious experience of the subject. This claim was based on the empirical finding that subjects sometimes fail to adopt optimal wagering strategies when asked to bet on their performance. We show here that when placed in a forced-choice situation and informed of chance level, subjects can perform above-chance in judging their own performance, in accordance with previous results from the literature (Kanai et al., 2010). Our findings therefore refute the hypothesis that no metacognitive knowledge can be accessed non-consciously and that confidence judgments or post-decision wagers could be good indices of awareness. Indeed, our remarks may be added to a list of other criticisms raised concerning the criterion shift induced by reward contingencies (Dienes and Scott, 2005; Fleming and Dolan, 2010; Schurger and Sher, 2008) in this measure, suggesting that indeed post-decision wagering might not constitute an ideal measure of awareness.

What then may be considered an appropriate measure of consciousness? This question is proving very difficult to answer. In particular, the problem of the ideal measure of consciousness depends tightly on the experimental question that is being addressed. In particular, when the goal of the experiment is to determine whether a specific process can be triggered non-consciously, it is crucial to assess with precision that the stimulus is indeed presented subliminally. In this case, it is commonly admitted that objective measure of performance assessed by *d'* provides the most stringent control of the absence of conscious perception, the use of subjective measures being considered superfluous. In the current thesis for instance, we used the shortest condition of SOA as a complete subliminal condition in which objective performance and detection rate were at chance, indicating total invisibility of the stimulus. Indeed, such an objective measure has been used in experiments that assessed the depth of non-conscious processing (Pessiglione et al., 2008; Pessiglione et al., 2008; Naccache et al., 2005). However, an important question for the field of consciousness is the impact of subjective conscious perception on cognitive processes, independently of the level of evidence that enters the system. In this respect, the question of whether a stimulus is "truly" subliminal might be a false problem. While cognitive scientists and philosophers agree that the study of consciousness is conscious experience itself and that this experience reflects a true state of the brain, subjective reports of perception constitute the true object of study when investigating the specificity of consciousness. In a way, this represents the only valid measure of conscious experience. Nonetheless, this does not mean that we should stop here. Indeed, while subjective measures offer an insight on the subjective experience, it is important for behavior to be characterized with as much precision as brain activity. In this respect, objective measures of performance, as well as detection abilities and confidence levels, should

be studied and documented with great detail. We believe such an approach of multiple measures of behavior would indeed help to characterize conscious content in its globality and understand precisely the characteristic of consciousness, rather than simply trusting one measure to assess it.

CHAPTER 8

Models of error-detection

8.1 Computational models of the ERN

We have seen that different computational models have been proposed for error detection and the ERN. Two main competing models of error-detection have proposed that the ERN behaves either as a mismatch or a conflict detector. Importantly, both models stipulate that the ERN comes from the confrontation of two signals: the representation of motor action and the representation of the correct/required response. However, they differ in the implementation of the computation of these responses as well as the nature of the confrontation process. Below, we highlight three key-points that diverge between these two models.

1. The two models differ in the underlying mathematical simulation regarding the amplitude of the ERN. While the comparison model supposes that the ERN represents a mismatch signal, corresponding to the subtraction between the representations of the correct and the actual response, conflict model however supposes that the ERN reflects the congruence between the two responses, in other word the product of these two signals.

2. The relationship between the computation of the required and the actual response is different in the two models: while the mismatch or comparison model supposes that the correct response and the actual response are computed independently (or at least no other possibility has been stipulated so far), the conflict model supposes that inhibitory connections exists between the two decisions units corresponding to the two representations. As a result, the computation of the actual and the required response constitutes in fact one single decisional process - the error resulting from initial incorrect activation that are further inhibited by the activation of the correct response. Importantly, the co-activation of the two responses can occur only very transiently, this state being highly unstable for the network.

3. Conflict theory supposes that conflict is assessed in a continuous fashion with the occurrence of a conflict signal not being time-locked to any particular neural event. Importantly, this model makes the prediction that on correct trials, conflict detection can occur at an early stage, prior to the response while only on error trials, conflict can be detected after the response. On the contrary, according to the mismatch theory, the comparison process is tightly locked to the motor response or alternatively, to the computation of the correct response (Falkenstein et al., 2000) with error-detection corresponding to a discrete and late event.

These three points each have an important impact on the models of the ERN and we discuss them in light of our findings. Regarding the first point, we can see from simulation data (Figure 8.1) that both models make slightly different predictions on the ERN amplitude. We tested these predictions when varying continuously the level of evidence on motor response and on the correct/required response from -1 (left response) to 1 (right response). For each model, we computed the mathematical solution proposed to reflect the ERN amplitude.

The results revealed that both models make very similar predictions when the level of evidence is high, both computations being able to predict if a trial is correct or erroneous with similar precision. Considering the mismatch model, we observed that the scale of the output values goes from 0 to 2. Maximal values are obtained when the two signals present the largest discrepancy. However, when both signals have identical level of evidence, even when very weak, the mismatch model predicts that the amplitude of the ERN should be constant. In particular, the ERN should have a similar amplitude when strong evidence is present in favour of both the required and the executed action and when no evidence is available on either signal. Importantly, a change of scale is needed for this model: the output value directly represents the absolute amplitude value of the ERN, rather than the direct microvolt measure of amplitude. A more appropriate measure would therefore be $-|m_i - i_i|$.

The conflict model however separates optimally correct and error trials when the level of evidence is maximal, while for lower level of evidence, the conflict measure is close to 0. Interestingly, this measure does not necessitate any change in the scaling: if an error is produced, a strong negative signal is emitted while if the response is correct, a positive signal is emitted. While these measures do not translate directly into microvolt's ERN amplitude, a shift in baseline could account for the difference in value.

Interestingly, we see that the most important difference between the predictions of the two models is when evidence is at its lowest, when no evidence is available either concerning the correct response or the motor response. When no information is available on the correct response, while conflict models predict that the ERN amplitude should remain very weak, the mismatch model makes the prediction that the amplitude should vary according to the amount of information corresponding to the motor activation (Figure 8.1). Such a finding seems to speak in favor of the conflict model as no clear evidence seems to exist that the level of motor activation influences the ERN (Rodríguez-fornells et al., 2002). Moreover, the conflict models seems to be quite consistent with data concerning the negativity after correct trials suggesting that when a larger amount of information is provided, the CRN is reduced while the ERN is increased, as we found in our data (Charles et al., 2013). Crucially, we found that the decoding pattern of results followed the prediction of the mathematical conflict measure (Charles et al., 2013), suggesting that indeed the product of activation or level of evidence might account better for error detection process than the subtraction.

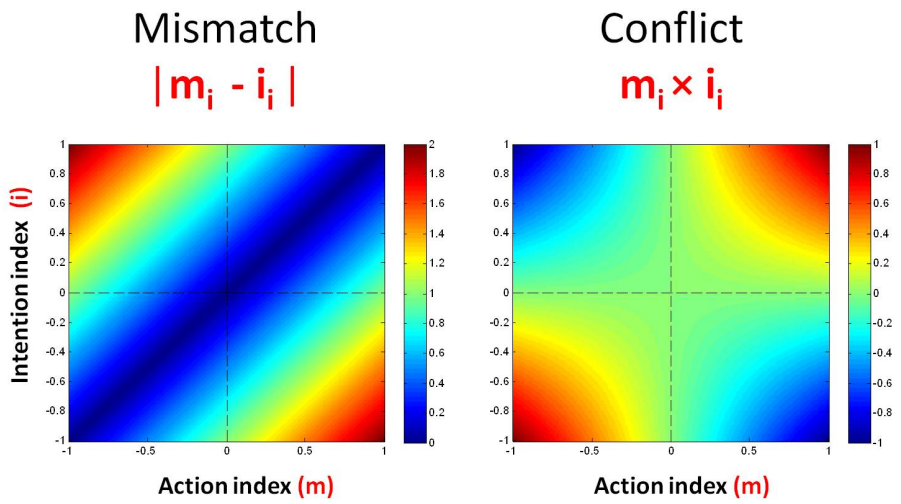


Figure 8.1: Simulation of the models of error-detection as a mismatch or as a conflict between intended and executed actions. In this simulation, we computed the prediction of both theories while varying the level of evidence on the actual motor response and the required response represented here as "motor index" and "intention index". We varied each value continuously from -1 (Left response) to +1 (Right Response) and we computed for each pair of values the mismatch (i.e. the absolute value of the difference) and the conflict (i.e. the product of the two values) predicted by each models. Blue colour corresponds to high evidence of error and red values correspond to high evidence for correct.

8.2 Dual versus single route model for decisions

Interestingly, however our data did not fit with the second or the third predictions of the conflict monitoring connectionist model. In particular, according to the conflict monitoring theory, both decision units accumulate evidence in favor of the response. Importantly however, as both units are linked by inhibitory connections, their patterns of activity are tightly linked together. Therefore, errors are characterized by the initial activation of the incorrect response unit followed rapidly by the activation of the correct response unit. In this respect, the conflict monitoring model is on the same line as single-route models of decisions in which errors results from fast motor activation. According to these models, incorrect responses do not make full use of all the available information from stimulus processing and are later followed by "change of minds", resulting in the correction of the initial erroneous response (Resulaj et al., 2009; Kiani and Shadlen, 2009).

Such a model seems difficult to reconcile with our data. In particular, as we trained our classifiers on both correct and error trials to identify a representation of the correct response, such a representation should be orthogonal to performance. However, single route models suggest that for error responses, the computation of the correct response comes only later, after the motor response. Indeed, detailed simulation of the conflict monitoring theory suggest that the evidence about correct-response starts to be available only after the response (Yeung et al., 2004) as can be seen on Figure 8.2.

However, in our present findings we found that it was possible to decode a representation of the

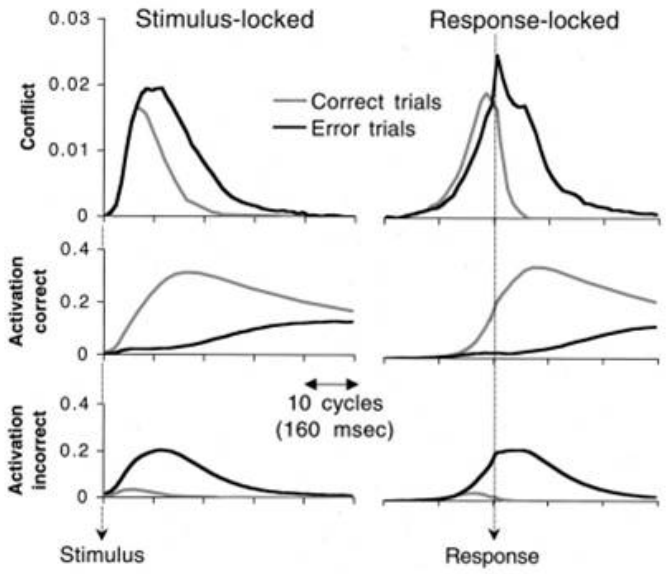


Figure 8.2: Simulations of the conflict monitoring model (from Yeung et al., 2004). Top graphs plots the simulated response conflict locked either to the stimulus onset (left graphs) or to the response onset (right graphs), for correct (grey line) and error trials (black line). The conflict value represents the product of the activity in the correct response unit (middle graphs) and the incorrect response unit (lower graphs). While correct trials are characterized by the massive activation of the correct response unit (middle graphs), error trials are characterized by the initial activation of the incorrect response unit (bottom graph) followed by the activation of the correct response unit (middle graphs).

correct response when training our decoder both on correct and error trials, suggesting that a common representation of the correct response existed in both types of trials. Importantly, we found that decoding was possible at an early stage of stimulus processing, suggesting that the computation of the correct response even on error trials was not limited to the time following the motor response. Moreover we found that the timing of the decoding of error detection was tightly locked with the timing of the computation of both the required and the actual response, occurring only at later stages - a finding that seem slightly at odds with the third prediction of conflict monitoring theory which would predict an earlier response.

To account for these findings, we proposed a dual-route model for error detection (Del Cul et al., 2009). According to this model, error detection would result from the comparison between the outputs of two distinct routes: a fast sensori-motor route that computes motor response and a slow but accurate conscious route that computes intentions. Importantly, whenever the output of one of those two routes diverges, an ERN is produced. Importantly, this models account also for the distinction between conscious and non-conscious perception, with only conscious conditions corresponding to the triggering of the higher conscious route and the emergence of a representation of the intended/required action.

Such a model is based on the classic accumulation or "random-walk" model for decision making (Ratcliff, 1985; Link, 2003; Laming, 1968; Ratcliff and Rouder, 1998). The decision system receives information about the stimulus in the form of noisy sensory inputs during a limited time-duration. The

evidence is integrated over time until the accumulated evidence in favor of one of the responses exceeds a fixed decision threshold. To model the masking paradigm, inputs can be presented for a certain duration that corresponds to the target-mask SOA, after which the system only receives noise. In the dual-route version of this model, conscious and non-conscious route have different characteristics (Figure 8.3):

1. The two parallel routes accumulates the same input of sensory evidence, but with different noise levels. A response can be emitted by whichever route first reaches its decision threshold.

2. The non-conscious route operates by continuous accumulation: a response can be either emitted when the threshold is reached or, if the threshold was not reached after a fixed duration, produce a response using the state of accumulated evidence at that moment.

3. The conscious route however operates only in an all-or-none mode. Importantly, the conscious route continues to accumulate evidence even if the lower route has reached its threshold. If the threshold is reached, the stimulus is labeled as *seen* . However, if after another time-delay the threshold is not reached, the trial is labeled as *unseen*.

The model predicts that while non-conscious information is still accumulated in the non-conscious route and can influence the motor-response, a stable representation of the correct response is triggered only when the conscious routes reached its threshold. Importantly, the two-routes accumulate evidence simultaneously, accounting for our finding that the correct response can be decoded in both correct and error trials. Furthermore, we predict that the ERN reflects the congruence of the outputs of these two routes. Importantly, the fast non-conscious route often produces fast-responses that are erroneous.

Interestingly, we believe such a model could also account for our patients' data showing that conscious mechanisms are impaired in schizophrenia while non-conscious mechanisms might be preserved. Indeed, the initial version of this model was intended to explain how prefrontal lesions can impact conscious report while maintaining the level of performance (Del Cul et al., 2009). As schizophrenia is associated with deficits in prefrontal activity (Barch et al., 2001), this model might be highly relevant to simulate patients' data regarding error-detection. For frontal patients, two possible variations of the model were originally tested: in one the threshold for conscious access was higher for patients than for controls while in the other the noise level in the conscious route was higher for patients than for controls. It was found that the second option provided a better fit of the data explaining that objective performance remained essentially unchanged when conditioned to subjective visibility. While more research is needed to understand if such a modification constitutes a proper model for schizophrenic patients, this model constitutes a plausible solution to explain specific deficits in conscious processing of schizophrenic patients.

This model is reminiscent of a previous account of cognitive control mechanisms, in which decision and motor control are organized in a hierarchical manner to monitor behavior (Norman and Shallice, 1980; Posner and Rothbart, 1998; Norman, 1981). Importantly, it follows the proposition that parallel decision processes could explain fast mechanisms of error correction and error compensation, following

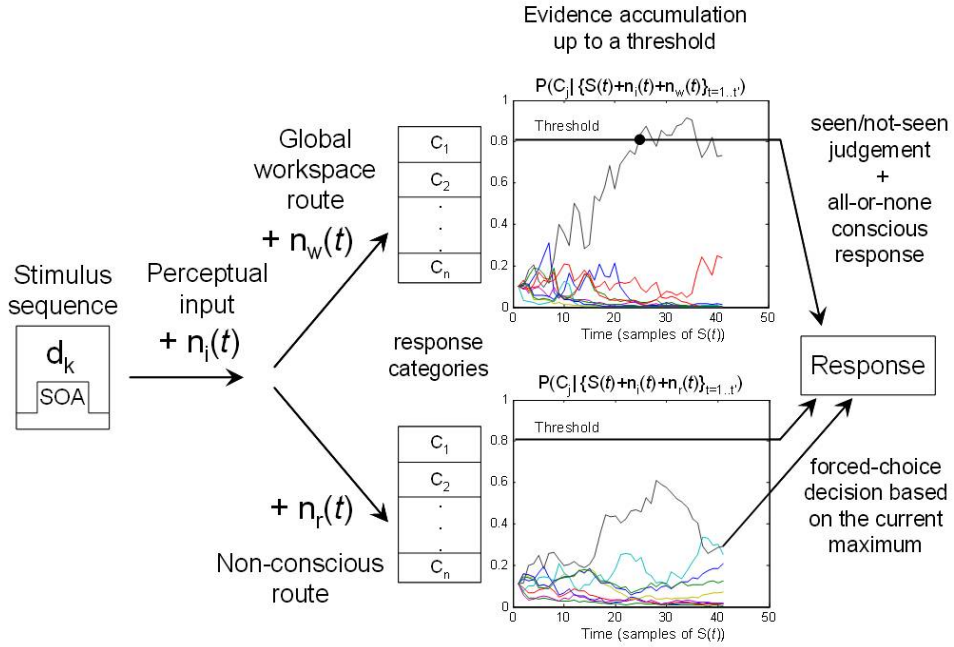


Figure 8.3: Dual-route model of conscious and non-conscious processing from Del Cul et al., 2009. See text.

suggestions of early studies on action slips and post-error adjustments (Norman, 1981; Rabbitt, 2002; Rabbitt and Vyas, 1981; Rabbitt, 1966b; Rabbitt, 1966a). While direct computational simulations are needed to test the details of this model, we believe it provides some interesting perspectives for the field of consciousness and error monitoring.

8.3 Are confidence judgments and error-detection processes the same?

As it is, the model provides a way to simulate how error might be detected. However, a question which remains is whether the model can account for confidence judgments as well. In particular, do confidence judgments and error detection correspond to the same processes?

Theoretical framework of second-order theory developed by Pleskac and Busemeyer (2010) provides a model of how confidence judgments and error detection might be linked. Indeed, according to these models, confidence judgments depend on the level of evidence that is reached after the initial crossing of the threshold, when continuing to integrate sensory information after a motor response have been emitted. According to this type of model that we can call "post-decisional locus model" (Yeung et al., 2004), confidence judgments consist in placing different criterion on the decision axis and making a confidence decision according to where the evidence falls (see Figure 1.13 on page 36). Interestingly, error-detection can be seen as a special case of this decisions in which instead of using several criteria distinguishing subtle levels of confidence, only one criterion is used to separate error from correct trials.

Such a modification would result in binary judgments of accuracy that would nonetheless be linked to confidence judgments.

Initial data on the ERN tends to validate this model. Scheffers and Coles (2000) showed that the ERN reflected confidence in the response. Could it be, then, that the ERN is an index of confidence judgment instead of an all-or-none error detection signal? In principle, such a view could be compatible with our model. Indeed we show that the amount of evidence on error-detection correlates with the amount of information regarding the required and the actual response. From a computational point of view, these two pieces of information might correspond respectively to the amount of evidence at the time of the motor response in the lower non-conscious route and the continuation of accumulation of evidence in the conscious-route after crossing its threshold. An important assumption that needs to be made however is that even for the conscious route, evidence continues to accumulate after crossing the threshold for conscious access. Indeed, without such a characteristic, our model, as well as any diffusion-to-bound model, makes the trivial prediction that all conscious content should be characterized by the same confidence level, corresponding to the threshold of conscious perception. As this assumption is not in the initial version of this model, precise simulations are needed to determine the validity of such an approach. It is possible to conceive that in a small amount of time after the crossing of the consciousness threshold, the content is refined and stabilized into a precise representation that contains more evidence than at the moment of the initial crossing of the threshold, this additional evidence serving as a basis for confidence judgments. However, no empirical evidence exists of such a mechanism at present.

Several criticisms can be raised on this model however. In a detailed article reviewing the question of confidence judgments versus error-detection, (Yeung and Summerfield, 2012) discusses these possible difficulties.

1. Models of confidence judgments that are based on level of evidence at decision time supposes that the observers can directly access these quantities (Pleskac and Busemeyer, 2010; Yeung and Summerfield, 2012). However, if evidence is directly available to the brain, it implies that a sampling process was not necessary in the first place, making the entire model meaningless.

2. While errors can often appear to be detected at a precise moment, confidence judgments often constitute a more continuous process, not particularly locked to a precise time. In particular, confidence judgments seem to be changing across time as we weight the difference source of evidence. This finding is difficult to reconcile with models of post-decisional locus as, according to these models, confidence indeed corresponds to a discrete decision process.

3. When reporting a confidence judgement, we often not only use the available evidence but also usually consider the trust we have in the source of the evidence. This view suggests that the level of evidence is not the only aspect that is taken into account for confidence judgments and that information regarding prior decisions and how they were distributed might also be involved.

Following part of these criticisms, Zylberberg et al. (2012) proposed a model of confidence in which

confidence judgments rely not on the locus of the evidence after the decision but rather on the decision-time, correlating with slope of the evidence accumulation process. Such a measure would reflect for each trial how "easy" the decision was to make. Provided that this measure is applied on the conscious route, which is based on a decision-to-bound model, such a view could explain the difference in confidence associated with different mental contents. However, such a measure does not address criticisms two and three as it supposes that confidence judgments correspond to discrete decision processes and more importantly, do not take into account distributions of prior decisions.

Yeung and Summerfield (2012) proposed an alternative model for confidence judgments taking into account the reliability of the evidence. They suggested that instead of only considering the mean of the strength of the decision variable, confidence judgment could also evaluate its variance. According to this view, the proper decision variable for confidence would be the probability distribution across the decision time (see Figure 1.15 on page 39). In this framework, the variance of the distribution of the decision process across time would provide an index of evidence reliability. Importantly, such a value could be computed in a continuous manner, providing a dynamic account of the evolution of confidence judgments (Yeung and Summerfield, 2012). Importantly, this measure is different from the one proposed by Zylberberg et al. (2012) as it accounts for the noise level in the accumulation process that might not properly addressed in the Zylberberg et al. (2012) model.

In light of these computational models of confidence, we suggest that confidence and error judgments might be distinct. While error judgments might be based on a classic diffusion-to-bound model assessing the congruence between the required and the actual response, confidence judgments might rely on a more complex statistical estimation, evaluating not only if the action matches the intention but also the reliability of the decision process itself. While more studies will be needed to test the validity of this approach, we believe the question of the relationship between confidence and error detection constitutes an important question to be addressed in the near future.

CHAPTER 9

Perspectives

After highlighting the possible convergence of our findings, we will discuss some points that remain to be studied and the future directions of research offered by the current work.

9.1 Action and Perception: the same status for consciousness?

While we have seen that consciousness and sensory information have a huge impact on the ERN, the relation of the ERN with motor action itself remains unclear. In particular, we showed that a conscious intention representing the required action was necessary for error detection and the ERN to be triggered, however several studies suggest that on the contrary, consciousness of the action is not a necessary factor for the ERN to occur. Using occulomotor tasks, Nieuwenhuis et al. (2001) showed that even when subjects failed to consciously detect the deviation in their own eye movements, brain activity is still characterized by the presence of an ERN, of identical amplitude. Similarly, Logan and Crump (2010) found that independently of conscious perception of accuracy, motor errors while typing words were still registered at some level and induced a noticeable impact on subsequent behavior. Indeed, pioneering work by Jeannerod (2003); Fourneret and Jeannerod (1998) suggested that we might have very little insight into our own motor actions and especially its fine modulation. Rather, we appear to monitor the goal of the action, leaving from conscious experience, the technical aspect of its execution.

Indeed, in the model of error detection that we propose, we suggest that motor actions may be triggered in majority by an unconscious route which is sensitive to priming and subliminal processing, while the conscious route might often arrive too late to directly influence motor output. Important points follow from this model. According to this view, motor action would constitute a non-conscious process registered only minimally by conscious experience that would simply index the conscious intention. Indeed, we would have only limited access to the details of our motor action or motor plans, detecting consciously only the outcome of actions.

Such a view remains to be tested. However, it constitutes a fundamental point for the research for consciousness. Indeed, the field of research of consciousness has concentrated its effort on understanding how consciousness relates to perceptual experience. However, much less evidence exists on the subject of subjective perception of our own acts. This question meets here the study of agency and self-awareness. However, before studying the question of how we take charge of our own actions, it is crucial to study precisely how much insight we have into our own action, in the same systematic manner that has been used to study conscious perception. Indeed, such study might lead us to approach the question

of the role of consciousness, keeping in mind that cognitive systems have primarily evolved to perform and control actions.

9.2 Metacognitive judgment of confidence outside of awareness

In this thesis, we found that even in non-conscious conditions where subjects deny consciously seeing the stimulus, they are nonetheless able to report their performance slightly better than chance. This finding, which has been suggested by previous results using the masking paradigm (Kanai et al., 2010), was replicated in three experiments, including one in schizophrenic patients. Crucially, our results in patients as well as neuroimaging data suggest that the metacognitive processes triggered non-consciously are distinct from the ones present in conscious conditions. Indeed patients presented a deficit in conscious metacognitive processes while non-conscious performance was identical to those of the controls. Furthermore, we found that distinct brain activity patterns were evoked in both conditions. In our initial source reconstruction of the M/EEG signal correlating with these above-chance reports, we found that the anterior region of the cingulate cortex seemed to be implicated, while more posterior activity was linked to error-detection in the conscious condition. However, this pattern of activity remained variable. It was found only when time-pressure was relaxed (Charles et al., 2013) and was not found to be as strong in the control group in our last experiment (Charles and Dehaene, 2013). This variability was further confirmed, as patients that presented similar above-chance estimation of their performance presented a very distinct pattern of brain activity, including more rostral regions of cingulate cortex as well as parahippocampic activity. Moreover, we were not able to train a decoder in the non-conscious condition to determine the accuracy of motor decisions, suggesting that the pattern of activity might be too variable to be decoded. Therefore, the neuronal substrate of this mechanism remains today slightly unclear and further research will be needed to understand exactly which brain regions encode this information.

An important question however is how such a metacognitive performance can occur outside of awareness. One first point that we have discussed is that such a mechanism corresponds to a forced-choice response rather than error detection per se. Indeed, to induce such a responses, subjects had to be informed that they had a fifty percent chance of responding correctly by chance. Pilot data suggested that when they did not receive such information, subjects tended to indicate that they had committed an error in non-conscious conditions. Therefore, these results might differ from mechanisms of all-or-none error detection and rather reflect a continuous statistical process of assessment of confidence in the response. Indeed, we have seen that confidence judgments might be based on mechanisms other than error detection, relying on the precision of evidence accumulation to produce a confidence judgment in the response. Importantly, such a mechanism might also be triggered non-consciously, providing an indirect measure of performance that might be predictive of response accuracy.

In our analysis of this effect, we excluded the trivial hypothesis that subjects were simply scanning their own reaction time to produce an accuracy judgment (see Supplementary information of Charles et al., 2013). However, many other aspects of decision might be relevant to predict performance, such as

scanning the variance in evidence-accumulation process, as well as the strength of motor activation. An important question is whether this process is truly non-conscious or whether it corresponds to a conscious amplification of non-conscious information. In particular, it has been proposed that conscious attentional mechanisms and task sets might have an effect on the processing of subliminal information. In this sense, the question remains of whether such a mechanism reflects a true conscious or non-conscious process. A related question that we could not address here specifically concerns the temporal dynamics of this process. In particular, we focused our analysis on the time following the response but this time-window, while relevant for the study of the ERN, might not correspond to those of this type of confidence judgments. Therefore, such a process still offers many interesting avenues to study that we hope will be addressed in the future.

CHAPTER 10

Conclusion

In this thesis, we investigated how metacognitive processes of error-detection relate to consciousness. We studied brain response to errors in conscious and subliminal conditions and showed that distinct brain processes are at stake in conscious and non-conscious conditions. We found that in conscious conditions, the brain computes a stable representation of the correct response, a conscious intention that codes for the required response associated to the stimulus. However, we proposed that sometimes this representation arrives too late to directly trigger motor action. Importantly, our results suggested that error detection and its underlying brain markers result from the comparison of these two brain signals, the congruence between these two representations signaling the correctness of the response. We showed that while this system of error-detection seems to be impeded in non-conscious conditions, some metacognitive judgments are still possible outside of awareness, indicating that statistical assessment of confidence in the response exists in non-conscious conditions. Importantly however this mechanism seems to be distinct from the one triggered in conscious conditions, as revealed by data from schizophrenic patients showing impaired conscious error-detection but preserved non-conscious meta-performance.

This work provides new findings regarding the depth of non-conscious processing that should have an impact on the field of consciousness research. In particular, our results suggest that metacognition is not the hallmark of consciousness and should not be used by itself as a measure of the level of consciousness. However, we showed that some processes vary in an all-or-none manner with consciousness, confirming that conscious access has a drastic impact on brain responses linked to performance monitoring. By isolating representations of abstract conscious decisions that are not directly related to the ongoing motor plan, we further extend the field to the question of the role of consciousness, offering novel avenues for investigation. Furthermore, this work provides the initial elements of a theoretical model of conscious and non-conscious processing that makes precise predictions on the dynamics of decisions and meta-decisions. This theoretical model constitutes a first attempt to bridge the gap between the fields of consciousness and metacognition, bringing novel insights to a promising field of research.

Bibliography

Aarts K, Pourtois G (2010) Anxiety not only increases, but also alters early error-monitoring functions. *Cognitive, Affective, & Behavioral Neuroscience* 10:479–492. (p 47.)

Agam Y, Hamalainen M, Lee ACH, Dyckman Ka, Friedman JS, Isom M, Makris N, Manoach DS (2011) Multimodal neuroimaging dissociates hemodynamic and electrophysiological correlates of error processing. *Proceedings of the National Academy of Sciences* 108:17556–17561. (p 47.)

Alain C (2002) Neurophysiological Evidence of Error-monitoring Deficits in Patients with Schizophrenia. *Cerebral Cortex* 12:840–846. (p 46, 62, 170.)

Aru J, Bachmann T, Singer W, Melloni L (2012) Distilling the neural correlates of consciousness. *Neuroscience & Biobehavioral Reviews* 36:737–746. (p 228.)

Azzopardi P, Cowey A (1998) Blindsight and visual awareness. *Consciousness and cognition* 7:292–311. (p 18.)

Baars BJ (1994) A Thoroughly Empirical Approach To Consciousness. *Psyche* 1. (p 7.)

Baars BJ (2002) The conscious access hypothesis: origins and recent evidence. *Trends in cognitive sciences* 6:47–52. (p 226.)

Barceló F, Escera C, Corral MJ, Perianez JA (2006) Task switching and novelty processing activate a common neural network for cognitive control. *J Cogn Neurosci* 18:1734–1748. (p 43.)

Barch DM, Carter CS, Braver TS, Sabb FW, MacDonald a, Noll DC, Cohen JD (2001) Selective deficits in prefrontal cortex function in medication-naive patients with schizophrenia. *Archives of general psychiatry* 58:280–288. (p 63, 235.)

Barch DM, Ceaser A (2011) Cognition in schizophrenia: core psychological and neural mechanisms. *Trends in Cognitive Sciences* 16:27–34. (p 63.)

Bassett DS, Bullmore E, Verchinski Ba, Mattay VS, Weinberger DR, Meyer-Lindenberg A (2008) Hierarchical organization of human cortical networks in health and schizophrenia. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 28:9239–9248. (p 64.)

Bates AT, Liddle PF, Kiehl Ka, Ngan ETC (2004) State dependent changes in error monitoring in schizophrenia. *Journal of psychiatric research* 38:347–356. (p 62, 170.)

Bates ATAT, Kiehl KAKa, Laurens KR, Liddle PFPF (2002) Error-related negativity and correct response negativity in schizophrenia. *Clinical Neurophysiology* 113:1454–1463. (p 45, 62, 170.)

Batterink L, Neville HJ (2013) The Human Brain Processes Syntax in the Absence of Conscious Awareness. *Journal of Neuroscience* 33:8528–8533. (p 19, 68.)

Becker GM, Degroot MH, Marschak J (1964) Measuring utility by a single-response sequential method. *Behavioral Science* 9:226–232. (p 26.)

Behrens TEJ, Woolrich MW, Walton ME, Rushworth MFS (2007) Learning the value of information in an uncertain world. *Nature neuroscience* 10:1214–1221. (p 28.)

Bogacz R, Brown E, Moehlis J, Holmes P, Cohen JD (2006) The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review* 113:700–765. (p 35.)

Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory* . (p 77.)

Botvinick M, Braver TS, Barch DM, Carter CS, Cohen JD (2001) Conflict monitoring and cognitive control. *Psychol Rev* 108:624–652. (p 42, 49.)

Botvinick M, Cohen JD, Carter CS (2004) Conflict monitoring and anterior cingulate cortex: an update. *Trends Cogn Sci* 8:539–546. (p 46, 49.)

Brazdil M, Roman R, Falkenstein M, Daniel P, Jurak P, Rektor I, Brázdil M (2002) Error processing–evidence from intracerebral ERP recordings. *Exp Brain Res* 146:460–466. (p 46, 47.)

Breitmeyer BG, Ogmen H (2006) *Visual masking: Time slices through conscious and unconscious vision*. (p 8.)

Brown JW, Braver TS (2005) Learned predictions of error likelihood in the anterior cingulate cortex. *Science* 307:1118–1121. (p 46.)

Bruchmann M, Herper K, Konrad C, Pantev C, Huster RJ (2010) Individualized EEG source reconstruction of Stroop interference with masked color words. *NeuroImage* 49:1800–1809. (p 8.)

Busch Na, Dubois J, VanRullen R (2009) The phase of ongoing EEG oscillations predicts visual perception. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 29:7869–7876. (p 226.)

Capa RL, Bustin GM, Cleeremans A, Hansenne M (2011) Conscious and unconscious reward cues can affect a critical component of executive control. *Experimental psychology* 58:370–375. (p 20, 223.)

Carter CS, Braver TS, Barch DM, Botvinick M, Noll D, Cohen JD (1998) Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science* 280:747–749. (p 49.)

Carter CS, MacDonald aW, Ross LL, Stenger Va (2001) Anterior cingulate cortex activity and impaired self-monitoring of performance in patients with schizophrenia: an event-related fMRI study. *The American journal of psychiatry* 158:1423–1428. (p 62, 63, 170.)

Chalmers DJDJ (1995) Facing up to the problem of consciousness. *Journal of Consciousness Studies* 2:20. (p 6.)

Charles B, Peirce S, Jastrow J (1885) On Small Differences in Sensation. *Memoirs of the National Academy of Sciences* . (p 24.)

Charles L, Dehaene S (2013) Preserved unconscious metacognition and impaired conscious error-detection in schizophrenia. *In preparation* . (p 84, 240.)

Charles L, King JR, Dehaene S (2013) Decoding the dynamics of action, intention, and error-detection for conscious and subliminal stimuli. *In revision* . (p 83, 232.)

Charles L, van Opstal F, Marti S, Dehaene S (2013) Distinct brain mechanisms for conscious versus subliminal error detection. *NeuroImage* 73:80–94. (p 83, 169, 170, 232, 240.)

Chevrier A, Schachar RJ (2010) Error detection in the stop signal task. *Neuroimage* 53. (p 46.)

Chiu PH, Deldin PJ (2007) Neural evidence for enhanced error detection in major depressive disorder. *The American journal of psychiatry* 164:608–616. (p 60.)

Chua EF, Schacter DL, Sperling Ra (2009) Neural correlates of metamemory: a comparison of feeling-of-knowing and retrospective confidence judgments. *Journal of cognitive neuroscience* 21:1751–1765. (p 27.)

Cohen A, Ivry RI, Keele SWSW (1990) Attention and structure in sequence learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* . (p 20.)

Cohen JD, Yeung N (2006) The impact of cognitive deficits on conflict monitoring. Predictable dissociations between the error-related negativity and N2. *Psychol Sci* 17:164–171. (p 50, 51.)

Cohen MX (2010) Error-related medial frontal theta activity predicts cingulate-related structural connectivity. *NeuroImage* 55:1373–1383. (p 46.)

Cohen MX, van Gaal S, Ridderinkhof KR, Lamme VaF (2009) Unconscious errors enhance prefrontal-occipital oscillatory synchrony. *Frontiers in human* 3:1–12. (p 20, 43, 67, 82, 223.)

Coles MGHMGH, Scheffers MKMK, Holroyd CB (2001) Why is there an ERN/Ne on correct trials? Response representations, stimulus-related components, and the theory of error-processing. *Biological psychology* 56:173–189. (p 48.)

Cortes C, Vapnik V (1995) Support-vector networks. *Machine Learning* 20:273–297. (p 78.)

Curran T, Keele SW (1993) Attentional and nonattentional forms of sequence learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* . (p 20.)

Curran T (1995) On The Neural Mechanisms of Sequence Learning . (p 20.)

Danielmeier C, Eichele T, Forstmann BU, Tittgemeyer M, Ullsperger M (2011) Posterior medial frontal cortex activity predicts post-error adaptations in task-related visual and motor areas. *The Journal of neuroscience* 31:1780–1789. (p 44.)

Danielmeier C, Ullsperger M (2011) Post-error adjustments. *Frontiers in psychology* 2:233. (p 41.)

Danion JM, Meulemans T, Kauffmann-Muller F, Vermaat H (2001) Intact implicit learning in schizophrenia. *The American journal of psychiatry* 158:944–948. (p 63, 169.)

Davis MH, Coleman MR, Absalom AR, Rodd JM, Johnsrude IS, Matta BF, Owen AM, Menon DK (2007) Dissociating speech perception and comprehension at reduced levels of awareness. *Proceedings of the National Academy of Sciences of the United States of America* 104:16032–16037. (p 19.)

De Pisapia N, Turatto M, Lin P, Jovicich J, Caramazza A (2011) Unconscious Priming Instructions Modulate Activity in Default and Executive Networks of the Human Brain. *Cerebral cortex (New York, N.Y. : 1991)* 22:639–649. (p 21, 223.)

Debener S, Ullsperger M, Siegel M, Fiehler K, Cramon DYV, Engel AK, von Cramon DY (2005) Trial-by-trial coupling of concurrent electroencephalogram and functional magnetic resonance imaging identifies the dynamics of performance monitoring. *J Neurosci* 25:11730–11737. (p 46, 47.)

Dehaene S, Artiges E, Naccache L, Martelli C, Viard A, Schurhoff F, Recasens C, Martinot ML, Leboyer M (2003) Conscious and subliminal conflicts in normal subjects and patients with schizophrenia: the role of the anterior cingulate. *Proceedings of the National Academy of Sciences of the United States of America* 100:13722. (p 63, 64, 169.)

Dehaene S, Changeux JP (2011) Experimental and theoretical approaches to conscious processing. *Neuron* 70:200–227. (p 64, 226, 227.)

Dehaene S, Changeux JP, Naccache L, Sackur J, Sergent C (2006) Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends in cognitive sciences* 10:204–211. (p 224, 225, 226, 227.)

Dehaene S, Naccache L (2001) Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition* 79:1–37. (p 226.)

Dehaene S, Naccache L, Le Clec'H G, Koechlin E, Mueller M, Dehaene-Lambertz G, van de Moortele PF, Le Bihan D (1998) Imaging unconscious semantic priming. *Nature* 395:597–600. (p 18.)

Dehaene S, Posner MI, Tucker DM (1994) Localization of a neural system for error detection and compensation. *Psychol. Sci.* 5:303–305. (p 44, 45, 46, 47.)

Del Cul A, Baillet S, Dehaene S (2007) Brain dynamics underlying the nonlinear threshold for access to consciousness. *PLoS Biol* 5:2408–2423. (p 68, 69, 87, 88.)

Del Cul A, Dehaene S, Leboyer M, Cul AD (2006) Preserved subliminal processing and impaired conscious access in schizophrenia. *Archives of general psychiatry* 63:1313. (p 63, 65, 68, 169.)

Del Cul A, Dehaene S, Reyes P, Bravo E, Slachevsky A (2009) Causal role of prefrontal cortex in the threshold for access to consciousness. *Brain* 132:2531–2540. (p 68, 234, 235, 236.)

Dennett DC (1993) Consciousness explained . (p 6.)

Dennett DC (1988) Quining Qualia. (p 6.)

Desender K, Van den Bussche E (2012) Is consciousness necessary for conflict adaptation? A state of the art. *Frontiers in human neuroscience* 6:3. (p 21.)

Desmurget M, Reilly KT, Richard N, Szathmari A, Mottolese C, Sirigu A (2009) Movement intention after parietal cortex stimulation in humans. *Science* 324:811–813. (p 228.)

Desmurget M, Sirigu A (2012) Conscious motor intention emerges in the inferior parietal lobule. *Current opinion in neurobiology* 22:1004–1011. (p 228.)

Destrebecqz a, Cleeremans a (2001) Can sequence learning be implicit? New evidence with the process dissociation procedure. *Psychonomic bulletin & review* 8:343–350. (p 20, 223.)

Dhar M, Wiersema JR, Pourtois G (2011) Cascade of Neural Events Leading from Error Commission to Subsequent Awareness Revealed Using EEG Source Imaging. *PLoS ONE* 6:e19578. (p 55.)

Diederich A (2006) Modeling the effects of payoff on response bias in a perceptual discrimination task : two-stage-processing hypothesis 68:194–207. (p 35.)

Dienes Z, Scott RB (2005) Measuring unconscious knowledge: distinguishing structural knowledge and judgment knowledge. *Psychological research* 69:338–351. (p 229.)

Dupoux E, de Gardelle V, Kouider S (2008) Subliminal speech perception and auditory streaming. *Cognition* 109:267–273. (p 19.)

Endrass T, Franke C, Kathmann N (2005) Error Awareness in a Saccade Countermanding Task. *Journal of Psychophysiology* 19:275–280. (p 55.)

Endrass T, Klawohn J, Preuss J, Kathmann N (2012) Temporospatial dissociation of Pe subcomponents for perceived and unperceived errors. *Frontiers in human neuroscience* 6:178. (p 55.)

Endrass T, Reuter B, Kathmann N (2007) ERP correlates of conscious error recognition: aware and unaware errors in an antisaccade task. *Eur J Neurosci* 26:1714–1720. (p 48, 55, 56, 67, 87, 223, 228.)

Enns JT, Di Lollo V (2000) What's new in visual masking? *Trends in cognitive sciences* 4:345–352. (p 8.)

Eriksen CW, O'Hara WP, Eriksen B (1982) Response competition effects in same-different judgments. *Perception & psychophysics* 32:261–270. (p 35.)

Eriksen CW, Coles MG, Morris LR, O'Hara WP (1985) An electromyographic examination of response competition. *Bulletin of the Psychonomic Society* 23. (p 35.)

Evans S, Azzopardi P (2007) Evaluation of a'bias-free'measure of awareness. *Spatial vision* 20:61–77. (p 30, 31, 224.)

Falkenstein M, Hohnsbein J, Hoormann J, Blanke L (1991) Effects of crossmodal divided attention on late ERP components. II. Error processing in choice reaction tasks. *Electroencephalography and clinical neurophysiology* 78:447–455. (p 44, 45, 48.)

Falkenstein M, Hoormann J, Christ S, Hohnsbein J (2000) ERP components on reaction errors and their functional significance: a tutorial. *Biol Psychol* 51:87–107. (p 44, 45, 46, 48, 51, 68, 123, 231.)

Falkenstein M, Hoormann J, Hohnsbein J (2001) Changes of error-related ERPs with age. *Experimental Brain Research* . (p 58.)

Fallgatter AJ, Herrmann MJ, Roemmler J, Ehlis AC, Wagener A, Heidrich A, Ortega G, Zeng Y, Lesch KP (2004) Allelic variation of serotonin transporter function modulates the brain electrical response for error processing. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology* 29:1506–1511. (p 61.)

Fiehler K, Ullsperger M, von Cramon DY (2005) Electrophysiological correlates of error correction. *Psychophysiology* 42:72–82. (p 46.)

Fitzgerald KD, Welsh RC, Gehring WJ, Abelson JL, Himle Ja, Liberzon I, Taylor SF (2005) Error-related hyperactivity of the anterior cingulate cortex in obsessive-compulsive disorder. *Biological psychiatry* 57:287–294. (p 60.)

Fleming SM, Dolan RJ (2010) Effects of loss aversion on post-decision wagering: Implications for measures of awareness. *Consciousness and cognition* 19:352–363. (p 229.)

Fleming SM, Weil RS, Nagy Z, Dolan RJ, Rees G (2010) Relating introspective accuracy to individual differences in brain structure. *Science* 329:1541–1543. (p 29, 30.)

Fletcher P, McKenna PJ, Friston KJ, Frith CD, Dolan RJ (1999) Abnormal cingulate modulation of fronto-temporal connectivity in schizophrenia. *NeuroImage* 9:337–342. (p 64.)

Fletcher PC, Henson RN (2001) Frontal lobes and human memory: insights from functional neuroimaging. *Brain : a journal of neurology* 124:849–881. (p 27.)

Foote AL, Crystal JD (2007) Metacognition in the rat. *Current biology : CB* 17:551–555. (p 26.)

Foti D, Kotov R, Bromet E, Hajcak G (2012) Beyond the Broken Error-Related Negativity: Functional and Diagnostic Correlates of Error Processing in Psychosis. *Biological psychiatry* 71:864–872. (p 62, 170.)

Fourneret P, Jeannerod M (1998) Limited conscious monitoring of motor performance in normal subjects. *Neuropsychologia* 36:1133–1140. (p 21, 239.)

Friston KJ (1998) The disconnection hypothesis. *Schizophrenia research* 30:115–125. (p 64.)

Friston KJ, Frith CD (1995) Schizophrenia: a disconnection syndrome. *Clin Neurosci* . (p 64.)

Friston KJ, Holmes AP, Poline JB (1995) Analysis of fMRI time-series revisited. *Neuroimage* . (p 77.)

Friston K (2005) Disconnection and cognitive dysmetria in schizophrenia. *The American journal of psychiatry* 162:429–432. (p 64.)

Frith CD, Frith U (2006) The neural basis of mentalizing. *Neuron* 50:531–534. (p 29.)

Gaillard R, Del Cul A, Naccache L, Vinckier F, Cohen L, Dehaene S (2006) Nonconscious semantic processing of emotional words modulates conscious access. *Proceedings of the National Academy of Sciences of the United States of America* 103:7524–7529. (p 19.)

Galvin SJ, Podd JV, Drga V, Whitmore J (2003) Type 2 tasks in the theory of signal detectability: discrimination between correct and incorrect decisions. *Psychonomic bulletin & review* 10:843–876. (p 31, 32, 34.)

Gehring WJ, Fencsik D (2001) Functions of the medial frontal cortex in the processing of conflict and errors. *Journal of Neuroscience* 21:9430. (p 45.)

Gehring WJ, Goss B, Coles MGH, Meyer DE, Donchin E (1993) A neural system for error detection and compensation. *Psychological Science* 4:385–390. (p 44, 45, 46, 48, 52.)

Gehring WJ, Himle J, Nisenson LG (2000) Action-monitoring dysfunction in obsessive-compulsive disorder. *Psychological science* pp. 1–6. (p 46, 60.)

Gigerenzer G, Hoffrage U, Kleinbölting H (1991) Probabilistic mental models: a Brunswikian theory of confidence. *Psychological review* 98:506–528. (p 25.)

Goodale MA, Pélisson D, Prablanc C (1986) Large adjustments in visually guided reaching do not depend on vision of the hand or perception of target displacement. *Nature* 320:748–750. (p 21.)

Gratton G, Coles MG, Sirevaag EJ, Eriksen CW, Donchin E (1988) Pre- and poststimulus activation of response channels: a psychophysiological analysis. *Journal of experimental psychology. Human perception and performance* 14:331–344. (p 35.)

Green MF, Nuechterlein KH, Breitmeyer BG, Mintz J (1999) Backward masking in unmedicated schizophrenic patients in psychotic remission: possible reflection of aberrant cortical oscillation. *The American journal of psychiatry* 156:1367–1373. (p 64.)

Greenwald AG, Draine SC, Abrams RL (1996) Three cognitive markers of unconscious semantic activation. *Science* 9221307. (p 18.)

Grillon ML, Oppenheim C, Varoquaux G, Charbonneau F, Devauchelle AD, Krebs MO, Baylé F, Thirion B, Huron C (2012) Hyperfrontality and hypoconnectivity during refreshing in schizophrenia. *Psychiatry research* 211:226–233. (p 64.)

Hajcak G, Foti D (2008) Errors Are Aversive Defensive Motivation and the Error-Related Negativity. *Psychological Science* 19:103–108. (p 60.)

Hajcak G, Franklin ME, Foa EB, Simons RF (2008) Increased error-related brain activity in pediatric obsessive-compulsive disorder before and after treatment. *The American journal of psychiatry* 165:116–123. (p 60.)

Hajcak G, McDonald N, Simons RF (2004) Error-related psychophysiology and negative affect. *Brain and cognition* 56:189–197. (p 60, 62, 170.)

Hajcak G, Simons RF (2008) Oops !.. I did it again : An ERP and behavioral study of double-errors. *Brain and Cognition* . (p 44.)

Hampton RR (2001) Rhesus monkeys know when they remember. *Proceedings of the National Academy of Sciences of the United States of America* 98:5359–5362. (p 26.)

Haraldsson H (2004) Transcranial Magnetic Stimulation in the investigation and treatment of schizophrenia: a review. *Schizophrenia Research* 71:1–16. (p 64.)

Hart JT (1965) Memory and the feeling-of-knowing experience. *Journal of educational psychology* 56:208–216. (p 22.)

Haynes JD, Driver J, Rees G (2005) Visibility reflects dynamic changes of effective connectivity between V1 and fusiform cortex. *Neuron* 46:811–821. (p 226.)

Haynes JD, Rees G (2005) Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature neuroscience* 8:686–691. (p 79.)

Herrmann MJ, Ro J, Ehlis Ac, Heidrich A, Fallgatter AJ, Rommler J (2004) Source localization (LORETA) of the error-related-negativity (ERN/Ne) and positivity (Pe). *Brain Res Cogn Brain Res* 20:294–299. (p 47.)

Hester R, Foxe JJ, Molholm S, Shpaner M, Garavan H (2005) Neural mechanisms involved in error processing: a comparison of errors made with and without awareness. *Neuroimage* 27:602–608. (p 55.)

Hewig J, Coles MGH, Trippe RH (2011) Dissociation of Pe and ERN/Ne in the conscious recognition of an error. *Psychophysiology* pp. 1–7. (p 54, 56.)

Hochman EY, Eviatar Z, Breznitz Z, Nevat M, Shaul S (2009) Source localization of error negativity: additional source for corrected errors. *Neuroreport* 20:1144–1148. (p 47.)

Hollard G, Massoni S, Vergnaud JC (2010) Subjective beliefs formation and elicitation rules: experimental evidence pp. 106–112. (p 26.)

Holmes AJ, Pizzagalli Da (2008) Spatiotemporal dynamics of error processing dysfunctions in major depressive disorder. *Archives of general psychiatry* 65:179–188. (p 60.)

Holroyd CB, Coles MGH (2002) The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychol Rev* 109:679–709. (p 52, 53.)

Holroyd CB, Dien J, Coles MGH (1998) Error-related scalp potentials elicited by hand and foot movements : evidence for an output-independent error-processing system in humans. *Brain* 242:65–68. (p 44, 45, 46.)

Holroyd CB, Krigolson OE, Baker R, Lee S, Gibson J (2009) When is an error not a prediction error? An electrophysiological investigation. *Cognitive, affective & behavioral neuroscience* 9:59–70. (p 52.)

Holroyd CB, Nieuwenhuis S, Yeung N, Cohen JD (2003) Errors in reward prediction are reflected in the event-related brain potential. *Neuroreport* 14:2481–2484. (p 52.)

Holt Ca, Smith AM (2009) An update on Bayesian updating. *Journal of Economic Behavior & Organization* 69:125–134. (p 26.)

Hughes G, Yeung N (2011) Dissociable correlates of response conflict and error awareness in error-related brain activity. *Neuropsychologia* 49:405–415. (p 48, 58.)

Jeannerod M (2003) Consciousness of action and self-consciousness. A cognitive neuroscience approach. *Agency and Self-Awareness: Issues in Philosophy and Psychology* pp. 128–149. (p 21, 239.)

Johannes S, Wieringa BM, Nager W, Rada D, Dengler R, Emrich HM, Münte TF, Dietrich DE (2001) Discrepant target detection and action monitoring in obsessive-compulsive disorder. *Psychiatry research* 108:101–110. (p 60.)

Kalayam B, Alexopoulos GS (2003) A preliminary study of left frontal region error negativity and symptom improvement in geriatric depression. *The American journal of psychiatry* 160:2054–2056. (p 60.)

Kanai R, Walsh V, Tseng CH (2010) Subjective discriminability of invisibility: A framework for distinguishing perceptual and attentional failures of awareness. *Consciousness and cognition* 19:1045–1057. (p 67, 223, 224, 229, 240.)

Kepecs A, Uchida N, Zariwala Ha, Mainen ZF (2008) Neural correlates, computation and behavioural impact of decision confidence. *Nature* 455:227–231. (p 26, 27.)

Kerns JG, Cohen JD, Macdonald AM, Johnson MK, Stenger VA, Aizenstein HJ, Carter CS (2005) Decreased conflict- and error-related activity in the anterior cingulate cortex in subjects with schizophrenia. *The American journal of psychiatry* 162:1833–1839. (p 62, 170.)

Kiani R, Shadlen MN (2009) Representation of confidence associated with a decision by neurons in the parietal cortex. *Science* 324:759–764. (p 26, 27, 28, 233.)

Kiehl KA, Liddle PF, Hopfinger JB (2000) Error processing and the rostral anterior cingulate: An event-related fMRI study. *Psychophysiology* . (p 49.)

Kim MS, Kang SS, Shin KS, Yoo SY, Kim YY, Kwon JS (2006) Neuropsychological correlates of error negativity and positivity in schizophrenia patients. *Psychiatry and clinical neurosciences* 60:303–311. (p 62, 170.)

Klein TA, Endrass T, Kathmann N, Neumann J, von Cramon DY, Ullsperger M (2007a) Neural correlates of error awareness. *Neuroimage* 34:1774–1781. (p 55.)

Klein TA, Neumann J, Reuter M, Hennig J, von Cramon DYY, Ullsperger M (2007b) Genetically determined differences in learning from errors. *Science* 318:1642. (p 44.)

Ko Y, Lau HC (2012) A detection theoretic explanation of blindsight suggests a link between conscious perception and metacognition. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 367:1401–1411. (p 34.)

Kolb FC, Braun J (1995) Blindsight in normal observers. *Nature* . (p 17, 67, 82.)

Kopp B, Rist F (1999) An event-related brain potential substrate of disturbed response monitoring in paranoid schizophrenic patients. *Journal of abnormal psychology* 108:337–346. (p 62, 170.)

Kouider S, de Gardelle V, Dehaene S, Dupoux E, Pallier C (2010) Cerebral bases of subliminal speech priming. *NeuroImage* 49:922–929. (p 19.)

Kouider S, Dehaene S (2007) Levels of processing during non-conscious perception: a critical review of visual masking. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 362:857–875. (p 8, 9.)

Kouider S, Dupoux E (2004) Partial awareness creates the "illusion" of subliminal semantic priming. *Psychological science* 15:75–81. (p 19.)

Kouider S, Dupoux E (2005) Subliminal speech priming. *Psychological science* 16:617–625. (p 19.)

Kriegeskorte N, Simmons WK, Bellgowan PSF, Baker CI (2009) Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience* 12:535–540. (p 76, 81.)

Kühn Aa, Williams D, Kupsch A, Limousin P, Hariz M, Schneider GH, Yarrow K, Brown P (2004) Event-related beta desynchronization in human subthalamic nucleus correlates with motor performance. *Brain : a journal of neurology* 127:735–746. (p 43.)

Kunimoto C, Miller J, Pashler HE (2001) Confidence and accuracy of near-threshold discrimination responses. *Consciousness and cognition* 10:294–340. (p 30, 229.)

Ladouceur CD, Dahl RE, Birmaher B, Axelson Da, Ryan ND (2006) Increased error-related negativity (ERN) in childhood anxiety disorders: ERP and source localization. *Journal of child psychology and psychiatry, and allied disciplines* 47:1073–1082. (p 60.)

Laming D (1979a) Autocorrelation of choice-reaction times. *Acta Psychologica* 43:381–412. (p 42.)

Laming D (1979b) Choice reaction performance following an error. *Acta Psychologica* 43:199–224. (p 42.)

Laming DRJ (1968) *Information theory of choice-reaction times.* (p 41, 42, 234.)

Lamme VaF, Roelfsema PR (2000) The distinct modes of vision offered by feedforward and recurrent processing. *Trends in neurosciences* 23:571–579. (p 20.)

Lau HC (2012) How to properly study the functions of consciousness? (p 226.)

Lau HC, Passingham RE (2006) Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proceedings of the National Academy of Sciences of the United States of America* 103:18763–18768. (p 17, 67, 82.)

Lau HC, Passingham RE (2007) Unconscious activation of the cognitive control system in the human prefrontal cortex. *J Neurosci* 27:5805–5811. (p 21, 67, 82, 223.)

Laurens KR (2003) Rostral anterior cingulate cortex dysfunction during error processing in schizophrenia. *Brain* 126:610–622. (p 62, 63, 170.)

Lemm S, Blankertz B, Dickhaus T, Müller KR (2011) Introduction to machine learning for brain imaging. *NeuroImage* 56:387–399. (p 81.)

Lesh Ta, Westphal AJ, Niendam Ta, Yoon JH, Minzenberg MJ, Ragland JD, Solomon M, Carter CS (2013) Proactive and reactive cognitive control and dorsolateral prefrontal cortex dysfunction in first episode schizophrenia. *NeuroImage: Clinical* . (p 63.)

Levine J (1983) Materialism and qualia: The explanatory gap. *Pacific philosophical quarterly* . (p 6.)

Liang M, Zhou Y, Jiang T, Liu Z, Tian L, Liu H, Hao Y (2006) Widespread functional disconnectivity in schizophrenia with resting-state functional magnetic resonance imaging. *Neuroreport* 17:209–213. (p 64.)

Link S (2003) C. S. Peirce, confidence and random walk theory. (p 234.)

Link SW (1975) The relative judgment theory of two choice response time. *Journal of Mathematical Psychology* . (p 35.)

Linkenkaer-Hansen K, Nikulin VV, Palva S, Ilmoniemi RJ, Palva JM (2004) Prestimulus oscillations enhance psychophysical performance in humans. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 24:10186–10190. (p 226.)

Logan GD, Crump MJC (2010) Cognitive illusions of authorship reveal hierarchical error detection in skilled typists. *Science* 330:683. (p 43, 60, 67, 223, 239.)

Luck SJ, Vogel EK, Shapiro KL (1996) Word meanings can be accessed but not reported during the attentional blink. *Nature* . (p 11, 18.)

Luu P, Flaisch T, Tucker DM (2000) Medial frontal cortex in action monitoring. *J Neurosci* 20:464–469. (p 45.)

Maier ME, Steinhauser M, Hubner R (2008) Is the error-related negativity amplitude related to error detectability? Evidence from effects of different error types. *Journal of Cognitive Neuroscience* 20:2263–2273. (p 56.)

Maier ME, Yeung N, Steinhauser M (2011) Error-related brain activity and adjustments of selective attention following errors. *NeuroImage* 56:2339–2347. (p 44.)

Mamassian P, Barthelme S (2009) Evaluation of Objective Uncertainty in the Visual System 5:1–8. (p 27.)

Maniscalco B, Lau HC (2012) A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and cognition* 21:422–430. (p 33, 34.)

Marco-Pallares J, Camara E, Munte TF, Rodríguez-Fornells A (2008) Neural mechanisms underlying adaptive actions after slips. *Journal of cognitive neuroscience* 20:1595–1610. (p 43, 44.)

Marois R, Chun MM, Gore JC (2000) Neural correlates of the attentional blink. *Neuron* 28:299–308. (p 11.)

Martens U, Ansorge U, Kiefer M (2011) Controlling the unconscious: attentional task sets modulate subliminal semantic and visuomotor processes differentially. *Psychological science* 22:282–291. (p 21, 223.)

Marti S, Sigman M, Dehaene S (2012) A shared cortical bottleneck underlying Attentional Blink and Psychological Refractory Period. *NeuroImage* 59:2883–2898. (p 11.)

Mathalon DH (2002) Response-monitoring dysfunction in schizophrenia: an event-related brain potential study. *Journal of abnormal psychology* . (p 62, 170.)

Mathalon DH, Bennett A, Askari N, Gray E, Rosenbloom MJ, Ford JM (2003) Response-monitoring dysfunction in aging and Alzheimer's disease: an event-related potential study. *Neurobiology of Aging* 24:675–685. (p 58.)

Mattler U (2003) Priming of mental operations by masked stimuli. *Perception & psychophysics* 65:167–187. (p 21, 223.)

McIntosh aR, Bookstein FL, Haxby JV, Grady CL (1996) Spatial pattern analysis of functional brain images using partial least squares. *NeuroImage* 3:143–157. (p 79.)

Metcalfe J, Glenberg A, Logan G, Nelson T, Oikawa L, Roediger H, Ryan P, Sharpe D, Steiger J (1986) Premonitions of Insight Predict Impending Error 12:623–634. (p 23, 24.)

Middlebrooks PG, Sommer MA (2012) Neuronal Correlates of Metacognition in Primate Frontal Cortex. *Neuron* 75:517–530. (p 27.)

Miltner WH, Lemke U, Weiss T, Holroyd CB, Scheffers MK, Coles MGH (2003) Implementation of error-processing in the human anterior cingulate cortex: a source analysis of the magnetic equivalent of the error-related negativity. *Biol Psychol* 64:157–166. (p 47.)

Modirrousta M, Fellows LK (2008) Dorsal medial prefrontal cortex plays a necessary role in rapid error prediction in humans. *The Journal of Neuroscience* 28:14000–14005. (p 27.)

Morris SE, Holroyd CB, Mann-Wrobel MC, Gold JM (2011) Dissociation of response and feedback negativity in schizophrenia: electrophysiological and computational evidence for a deficit in the representation of value. *Frontiers in human neuroscience* 5:123. (p 62, 170.)

Morris SE, Yee CM, Nuechterlein KH (2006) Electrophysiological analysis of error monitoring in schizophrenia. *Journal of abnormal psychology* 115:239–250. (p 62, 170.)

Moser JS, Hajcak G, Simons RF (2005) The effects of fear on performance monitoring and attentional allocation. *Psychophysiology* 42:261–268. (p 60.)

Munro GES, Dywan J, Harris GT, McKee S, Unsal A, Segalowitz SJ (2007) ERN varies with degree of psychopathy in an emotion discrimination task. *Biological psychology* 76:31–42. (p 46.)

Naccache L, Dehaene S (2001) Unconscious semantic priming extends to novel unseen stimuli. *Cognition* 80:215–229. (p 18.)

Naccache L, Gaillard R, Adam C, Hasboun D, Clémenceau S, Baulac M, Dehaene S, Cohen L (2005) A direct intracranial record of emotions evoked by subliminal words. *Proceedings of the National Academy of Sciences of the United States of America* 102:7713–7717. (p 19, 68, 229.)

Nagel T (1974) What is it like to be a bat? *The philosophical review* . (p 6.)

Nandy AS, Tjan BS (2012) Saccade-confounded image statistics explain visual crowding. *Nature neuroscience* 15:463–469. (p 10.)

Nelson TO, Narens L (1990) Metamemory: A theoretical framework and new findings. *The psychology of learning and motivation* 26. (p 22.)

Nieuwenhuis S, Ridderinkhof KR, Blom JH, Band GPH, Kok A (2001) Error-related brain potentials are differentially related to awareness of response errors: evidence from an antisaccade task. *Psychophysiology* 38:752–760. (p 43, 55, 67, 83, 87, 223, 228, 239.)

Nieuwenhuis S, Ridderinkhof KR, Talsma D, Coles MGH, Holroyd CB, Kok A, van der Molen MW (2002) A computational account of altered error processing in older age: dopamine and the error-related negativity. *Cognitive, affective & behavioral neuroscience* 2:19–36. (p 58.)

Norman DA (1981) Categorization of Action Slips. *Psychological Review* 88:1–15. (p 235, 236.)

Norman DA, Shallice T (1980) Attention to action: Willed and automatic control of behavior. *R. J. Davidson, G. E. Schwarts, & D. Shapiro (Eds.). Consciousness of self-regulation. Advances in research and theory (Vol. 4). New York: Plenum Press* pp. 1–18. (p 235.)

Norman Ka, Polyn SM, Detre GJ, Haxby JV (2006) Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in cognitive sciences* 10:424–430. (p 79.)

Notebaert W, Houtman F, van Opstal F, Gevers W, Fias W, Verguts T (2009) Post-error slowing: An orienting account. *Cognition* 111:275–279. (p 41, 43.)

Núñez Castellar E, Kühn S, Fias W, Notebaert W (2010) Outcome expectancy and not accuracy determines posterror slowing: ERP support. *Cognitive, affective and behavioral neuroscience* 10:270–278. (p 41.)

O'Connell RG, Bellgrove MA, Dockree PM, Lau A, Hester R, Garavan H, Fitzgerald M, Foxe JJ, Robertson IH (2009) The neural correlates of deficient error awareness in attention-deficit hyperactivity disorder (ADHD). *Neuropsychologia* 47:1149–1159. (p 55.)

O'Connell RG, Dockree PM, Bellgrove MA, Kelly SP, Hester R, Garavan H, Robertson IH, Foxe JJ (2007) The role of cingulate cortex in the detection of errors with and without awareness: a high-density electrical mapping study. *Eur J Neurosci* 25:2571–2579. (p 46, 48, 55.)

Olvet DMMDM, Hajcak G (2008) The error-related negativity (ERN) and psychopathology: Toward an endophenotype. *Clinical psychology review* 28:1343–1354. (p 58, 60, 62, 170.)

Orr JM, Carrasco M (2011) The role of the error positivity in the conscious perception of errors. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 31:5891–5892. (p 56.)

Overgaard M, Timmermans B, Sandberg K, Cleeremans A (2010) Optimizing subjective measures of consciousness. *Consciousness and cognition* 19:682–686. (p 16.)

Pailing PE, Segalowitz SJ (2004a) The effects of uncertainty in error monitoring on associated ERPs. *Brain and cognition* 56:215–233. (p 46, 48, 52, 87.)

Pailing PE, Segalowitz SJ (2004b) The error-related negativity as a state and trait measure: motivation, personality, and ERPs in response to errors. *Psychophysiology* 41:84–95. (p 62, 170.)

Pailing PE, Segalowitz SJ, Dywan J, Davies PL (2002) Error negativity and response control. *Psychophysiology* 39:198–206. (p 46.)

Passingham RE, Bengtsson SL, Lau HC (2010) Medial frontal cortex: from self-generated action to reflection on one's own performance. *Trends in cognitive sciences* 14:16–21. (p 29.)

Pavone EFEF, Marzi CA, Girelli M (2009) Does subliminal visual perception have an error-monitoring system? *Eur J Neurosci* 30:1424–1431. (p 52, 56, 68, 83, 87.)

Persaud N, McLeod P, Cowey A (2007) Post-decision wagering objectively measures awareness. *Nature Neuroscience* 10:257–261. (p 25, 26, 67, 82, 87, 229.)

Pessiglione M, Petrovic P, Daunizeau J, Palminteri S, Dolan RJ, Frith CD (2008) Subliminal instrumental conditioning demonstrated in the human brain. *Neuron* 59:561–567. (p 20, 67, 68, 82, 223, 229.)

Pessiglione M, Schmidt L, Draganski B, Kalisch R, Lau HC, Dolan RJ, Frith CD (2007) How the brain translates money into force: a neuroimaging study of subliminal motivation. *Science* 316:904. (p 20, 67, 68, 82, 87, 223.)

Platt JC (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* 10:61—-74. (p 79.)

Pleskac TJ, Busemeyer JR (2010) Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychological review* 117:864–901. (p 35, 36, 38, 68, 236, 237.)

Posner MI, Rothbart MK (1998) Attention, self-regulation and consciousness. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 353:1915–1927. (p 235.)

Qiao E, Vinckier F, Szwed M, Naccache L, Valabrègue R, Dehaene S, Cohen L (2010) Unconsciously deciphering handwriting: subliminal invariance for handwritten words in the visual word form area. *NeuroImage* 49:1786–1799. (p 19.)

Rabbitt PMA (1966a) Error correction time without external error signals. *Nature* 212:438. (p 35, 41, 236.)

Rabbitt PMA (1966b) Errors and error correction in choice-response tasks. *Journal of Experimental Psychology* 71:264–272. (p 35, 41, 236.)

Rabbitt PMA (2002) Consciousness is slower than you think. *Q J Exp Psychol A* 55:1081–1092. (p 35, 37, 56, 236.)

Rabbitt PMA, Vyas S (1981) Processing a display even after you make a response to it: How perceptual errors can be corrected. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology* 33. (p 36, 236.)

Rahnev Da, Bahdo L, de Lange FP, Lau HC (2012a) Prestimulus hemodynamic activity in dorsal attention network is negatively associated with decision confidence in visual perception. *Journal of neurophysiology* 108:1529–1536. (p 34.)

Rahnev Da, Maniscalco B, Luber B, Lau HC, Lisanby SH (2012b) Direct injection of noise to the visual cortex decreases accuracy but increases decision confidence. *Journal of neurophysiology* 107:1556–1563. (p 34.)

Ramsø y TZ, Overgaard M (2004) Introspection and subliminal perception. *Phenomenology and the Cognitive Sciences* pp. 1–23. (p 16.)

Ratcliff R (1985) Theoretical interpretations of the speed and accuracy of positive and negative responses. *Psychological review* 92:212–225. (p 35, 234.)

Ratcliff R (1978) A theory of memory retrieval. *Psychological review* . (p 35.)

Ratcliff R, Mckoon G (2009) The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks 20:1–44. (p 35.)

Ratcliff R, Rouder JN (1998) Modeling Response Times for Two-Choice Decisions. *Psychological Science* 9:347–356. (p 234.)

Reber TP, Luechinger R, Boesiger P, Henke K (2012) Unconscious relational inference recruits the hippocampus. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 32:6138–6148. (p 20.)

Reed J, Johnson P (1994) Assessing implicit learning with indirect tests: Determining what is learned about sequence structure. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20. (p 20.)

Resulaj A, Kiani R, Wolpert DM, Shadlen MN (2009) Changes of mind in decision-making. *Nature* 461:263–266. (p 36, 37, 38, 233.)

Reuss H, Kiesel A, Kunde W, Hommel B (2011) Unconscious activation of task sets. *Consciousness and cognition* 20:556–567. (p 21, 223.)

Reynvoet B, Ratinckx E (2004) Hemispheric differences between left and right number representations: effects of conscious and unconscious priming. *Neuropsychologia* 42:713–726. (p 19.)

Ridderinkhof KR (2002) Micro- and macro-adjustments of task set: activation and suppression in conflict tasks. *Psychological research* 66:312–323. (p 43.)

Riesel A, Weinberg A, Endrass T, Meyer A, Hajcak G (2013) The ERN is the ERN is the ERN? Convergent validity of error-related brain activity across different tasks. *Biological psychology* pp. 1–9. (p 45.)

Rodríguez-fornells A, Kurzbuch AR, Münte TF (2002) Time course of error detection and correction in humans: neurophysiological evidence. *The Journal of neuroscience* 22:9990–9996. (p 46, 232.)

Rounis E, Maniscalco B, Rothwell JC, Passingham RE, Lau HC (2010) Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. *Cognitive Neuroscience* 1:165–175. (p 17, 29, 34, 67, 82, 87.)

Saccuzzo DS, Cadenhead KS, Braff DL (1996) Backward versus forward visual masking deficits in schizophrenic patients: centrally, not peripherally, mediated? *The American journal of psychiatry* 153:1564–1570. (p 64.)

Sadaghiani S, Hesselmann G, Kleinschmidt A (2009) Distributed and antagonistic contributions of ongoing activity fluctuations to auditory stimulus detection. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 29:13410–13417. (p 19.)

Sandberg K, Timmermans B, Overgaard M, Cleeremans A (2010) Measuring consciousness: is one measure better than the other? *Consciousness and cognition* 19:1069–1078. (p 16.)

Scheffers MK, Coles MG, Bernstein P, Gehring WJ, Donchin E (1996) Event-related brain potentials and error-related processing: an analysis of incorrect responses to go and no-go stimuli. *Psychophysiology* 33:42–53. (p 48.)

Scheffers MK, Coles MGH (2000) Performance monitoring in a confusing world: error-related brain activity, judgments of response accuracy, and types of errors. *J Exp Psychol Hum Percept Perform* 26:141–151. (p 46, 48, 53, 54, 56, 237.)

Schie HTV, Mars RB, Coles MGH, Bekkering H, van Schie HT (2004) Modulation of activity in medial frontal and motor cortices during error observation. *Nature Neuroscience* 7:549–554. (p 44.)

Schmidt L, Palminteri S, Lafargue G, Pessiglione M (2010) Splitting motivation: unilateral effects of subliminal incentives. *Psychological science* 21:977–983. (p 20, 223.)

Schmitt A, Hasan A, Gruber O, Falkai P (2011) Schizophrenia as a disorder of disconnectivity. *European archives of psychiatry and clinical neuroscience* 261 Suppl:S150—-4. (p 64.)

Schurger A, Sher S (2008) Awareness, loss aversion, and post-decision wagering. *Trends in cognitive sciences* 12:209—-10; author reply 210. (p 229.)

Searle JR (1998) How to study consciousness scientifically. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 353:1935–1942. (p 7.)

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. (p 225.)

Sergent C, Dehaene S (2004a) Is consciousness a gradual phenomenon? Evidence for an all-or-none bifurcation during the attentional blink. *Psychol Sci* 15:720–728. (p 16, 225.)

Sergent C, Dehaene S (2004b) Neural processes underlying conscious perception: experimental findings and a global neuronal workspace framework. *J Physiol Paris* 98:374–384. (p 225, 226.)

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. *Current Biology* 23:150–155. (p 22.)

Shalgi S, Barkan I, Deouell LY (2009) On the positive side of error processing: error-awareness positivity revisited. *The European journal of neuroscience* 29:1522–1532. (p 55.)

Shalgi S, Deouell LY (2012) Is any awareness necessary for an Ne? *Frontiers in human neuroscience* 6:1–15. (p 54, 56.)

Sklar AY, Levy N, Goldstein A, Mandel R, Maril A, Hassin RR (2012) Reading and doing arithmetic nonconsciously 2012. (p 19, 68.)

Smith EE, Eich TS, Cebenoyan D, Malapani C (2011) Intact and impaired cognitive-control processes in schizophrenia. *Schizophrenia research* 126:132–137. (p 63.)

Smith JD, Shields WE, Washburn Da (2003) The comparative psychology of uncertainty monitoring and metacognition. *The Behavioral and brain sciences* 26:317–373. (p 26.)

Sperling G (1960) The information available in brief visual presentations. *Psychological monographs: General and applied* . (p 22, 23.)

Steele JD, Meyer M, Ebmeier KP (2004) Neural predictive error signal correlates with depressive illness severity in a game paradigm. *NeuroImage* 23:269–280. (p 60.)

Steinhauser M, Yeung N (2010) Decision processes in human performance monitoring. *The Journal of Neuroscience* 30:15643–15653. (p 56, 228.)

Stemmer B, Segalowitz SJ, Dywan J, Panisset M, Melmed C (2007) The error negativity in nonmedicated and medicated patients with Parkinson's disease. *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology* 118:1223–1229. (p 60.)

Strozyk JV, Jentzsch I (2012) Weaker error signals do not reduce the effectiveness of post-error adjustments: Comparing error processing in young and middle-aged adults. *Brain research* 1460C:41–49. (p 41.)

Swick D, Turken, U (2002) Dissociation between conflict detection and error monitoring in the human anterior cingulate cortex. *Proceedings of the National Academy of Sciences* 99:16354–16359. (p 60, 61, 62.)

Szczepanowski R, Pessoa L (2007) Fear perception: Can objective and subjective awareness measures be dissociated? *Journal of Vision* 7:1–17. (p 17.)

Todd MT, Nystrom LE, Cohen JD (2013) Confounds in Multivariate Pattern Analysis: Theory and Rule Representation Case Study. *NeuroImage* . (p 81.)

Tononi G, Edelman GM (2000) Schizophrenia and the mechanisms of conscious integration. *Brain research. Brain research reviews* 31:391–400. (p 169.)

Tsuchiya N, Koch C (2005) Continuous flash suppression reduces negative afterimages. *Nature neuroscience* 8:1096–1101. (p 10.)

Tsujimoto S, Genovesio A, Wise SP (2010) Evaluating self-generated decisions in frontal pole cortex of monkeys. *Nature neuroscience* 13:120–126. (p 27.)

Turken AU, Swick D (2008) The effect of orbitofrontal lesions on the error-related negativity. *Neuroscience Letters* 441:7–10. (p 60.)

Uhlhaas PJ, Haenschel C, Nikolic D, Singer W (2008) The role of oscillations and synchrony in cortical networks and their putative relevance for the pathophysiology of schizophrenia. *Schizophrenia bulletin* 34:927–943. (p 64.)

Ullsperger M (2006) Performance monitoring in neurological and psychiatric patients. *International journal of psychophysiology : official journal of the International Organization of Psychophysiology* 59:59–69. (p 61.)

Ullsperger M, Cramon DYV, von Cramon DY (2003) Error monitoring using external feedback: specific roles of the habenular complex, the reward system, and the cingulate motor area revealed by functional magnetic resonance imaging. *J Neurosci* 23:4308–4314. (p 47.)

Ulrich R, Szymanowski F (2004) ERP Correlates of error relevance. *Errors, conflicts, and the Brain.* . (p 43.)

Van den Bussche E, Notebaert K, Reynvoet B (2009) Masked primes can be genuinely semantically processed: a picture prime study. *Experimental psychology* 56:295–300. (p 19, 68.)

van Gaal S, de Lange FP, Cohen MX (2012) The role of consciousness in cognitive control and decision making. *Frontiers in human neuroscience* 6:121. (p 21.)

van Gaal S, Ridderinkhof KR, Fahrenfort JJ, Scholte HS, Lamme VaF (2008) Frontal cortex mediates unconsciously triggered inhibitory control. *J Neurosci* 28:8053–8062. (p 20, 21, 67, 82, 87, 223.)

van Gaal S, Ridderinkhof KR, Scholte HS, Lamme VaF (2010) Unconscious activation of the prefrontal no-go network. *The Journal of Neuroscience* 30:4143. (p 20, 21, 223.)

van Gaal S, Ridderinkhof KR, van den Wildenberg WPM, Lamme VaF (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. (p 20, 21, 67, 82, 87, 223.)

Van Opstal F, de Lange FP, Dehaene S (2011) Rapid parallel semantic processing of numbers without awareness. *Cognition* 120:136–147. (p 19.)

van Veen V, Carter CS (2006) Error detection, correction, and prevention in the brain: a brief review of data and theories. *Clin EEG Neurosci* 37:330–335. (p 41, 48.)

Van Veen V, Carter CS (2002) The timing of action-monitoring processes in the anterior cingulate cortex. *Journal of cognitive neuroscience* 14:593–602. (p 46, 49, 50, 51.)

Veen VV, Carter CS (2002) The anterior cingulate as a conflict monitor: fMRI and ERP studies. *Physiol Behav* 77:477–482. (p 46, 48, 49, 51.)

Veen VV, Cohen JD, Botvinick M, Stenger VA, Carter CS, van Veen V (2001) Anterior cingulate cortex, conflict monitoring, and levels of processing. *Neuroimage* 14:1302–1308. (p 49.)

Vidal F, Hasbroucq T, Grapperon J, Bonnet M (2000) Is the 'error negativity' specific to errors? *Biol Psychol* 51:109–128. (p 45.)

Vlamings P (2008) Reduced error monitoring in children with autism spectrum disorder: an ERP study. *European Journal of Neuroscience* 28:399–406. (p 46.)

Vocat R, Pourtois G, Vuilleumier P (2008) Unavoidable errors: a spatio-temporal analysis of time-course and neural sources of evoked potentials associated with error processing in a speeded task. *Neuropsychologia* 46:2545–2555. (p 46, 48.)

Voss A, Rothermund K, Voss J (2004) Interpreting the parameters of the diffusion model: an empirical validation. *Memory & cognition* 32:1206–1220. (p 35.)

Wagner AD (1998) Building Memories: Remembering and Forgetting of Verbal Experiences as Predicted by Brain Activity. *Science* 281:1188–1191. (p 27.)

Weibel S, Giersch A, Dehaene S, Huron C (2013) Unconscious task set priming with phonological and semantic tasks. *Consciousness and cognition* 22:517–527. (p 19.)

Weinberg A, Olvet DM, Hajcak G (2010) Increased error-related brain activity in generalized anxiety disorder. *Biological psychology* 85:472–480. (p 60.)

Weiskrantz L (1986) Blindsight: A Case Study and Implications . (p 18.)

Weiskrantz L (1996) Blindsight revisited. *Current opinion in neurobiology* 6:215–220. (p 17, 18.)

Wessel JR (2012) Error awareness and the error-related negativity: evaluating the first decade of evidence. *Frontiers in human neuroscience* 6:88. (p 52, 55, 59.)

Wessel JR, Danielmeier C, Ullsperger M (2011) Error awareness revisited: accumulation of multimodal evidence from central and autonomic nervous systems. *Journal of Cognitive Neuroscience* 23:3021–3036. (p 43, 56.)

West R, Travers S Tracking the Temporal Dynamics of Updating Cognitive Control: An Examination of Error Processing. *Cerebral cortex* 18:1112–1124. (p 45.)

Woodman GFF (2010) Masked targets trigger event-related potentials indexing shifts of attention but not error detection. *Psychophysiology* 47:410–414. (p 52, 57, 58, 68, 83, 87.)

Wyart V, Sergent C (2009) The phase of ongoing EEG oscillations uncovers the fine temporal structure of conscious perception. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 29:12839–12841. (p 226.)

Yeh SL, He S, Cavanagh P (2012) Semantic priming from crowded words. *Psychological science* 23:608–616. (p 19.)

Yeung N, Botvinick M, Cohen JD (2004) The Neural Basis of Error Detection: Conflict Monitoring and the Error-Related Negativity. *Psychological Review* 111:931–959. (p 41, 49, 50, 51, 68, 123, 233, 234, 236.)

Yeung N, Summerfield C (2012) Metacognition in human decision-making: confidence and error monitoring. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 367:1310–1321. (p 38, 39, 237, 238.)

Yoshida W, Ishii S (2006) Resolution of uncertainty in prefrontal cortex. *Neuron* 50:781–789. (p 29.)

Zetzsche T, Preuss U, Frodl T, Watz D, Schmitt G, Koutsouleris N, Born C, Reiser M, Möller HJ, Meisenzahl EM (2007) In-vivo topography of structural alterations of the anterior cingulate in patients with schizophrenia: new findings and comparison with the literature. *Schizophrenia research* 96:34–45. (p 63.)

Zhou FA, Davis G (2012) Unconscious priming of task sets: the role of spatial attention. *Attention, perception & psychophysics* 74:105–114. (p 21, 223.)

Zylberberg A, Barttfeld P, Sigman M, Pereira A (2012) The construction of confidence in a perceptual decision. *Frontiers in integrative neuroscience* 6:79. (p 38, 237, 238.)

Zylberberg A, Fernández Slezak D, Roelfsema PR, Dehaene S, Sigman M (2010) The brain's router: a cortical network model of serial processing in the primate brain. *PLoS computational biology* 6:e1000765. (p 11.)

Lucie CHARLES

MECANISMES CONSCIENTS ET NON-CONSCIENTS DE LA DECISION ET DE LA « META-DECISION »

Dans cette thèse, nous avons étudié les liens entre conscience et metacognition. Les processus métacognitifs correspondent à l'évaluation, au contrôle et à l'introspection de notre propre cognition. Souvent considérés comme le propre de la conscience, cette association doit néanmoins être testée expérimentalement. Nous nous sommes concentrés sur la détection d'erreur afin d'étudier comment l'expérience subjective consciente influence les réponses cérébrales magnéto-et électro-encéphalographiques (M/EEG) liées au contrôle de la performance. Dans une première étude, nous avons montré que l'ERN, un marqueur cérébral de la détection d'erreur est absent dans des conditions subliminales alors que les sujets sont encore en mesure de prédire leur performance mieux que le hasard. Ces résultats suggèrent que deux processus de contrôle de la performance coexistent: l'ERN, un signal tout-ou-rien de détection d'erreur, est présent uniquement dans des conditions conscientes, alors que la confiance dans la réponse peut être estimée dans des conditions non-conscientes. Pour tester si ces deux processus sont véritablement distincts, nous avons reproduit notre étude chez une population de patients schizophrènes qui présentent des déficits spécifiques dans les conditions de perception consciente. Nos résultats montrent que les patients ont des performances métacognitives normales dans des conditions subliminales, alors que les processus de détection d'erreurs conscients sont altérés, ce qui confirme la distinction entre détection d'erreur consciente et non-consciente. Pour étudier plus précisément la nature de cette différence, nous avons utilisé les méthodes de décodage appliquées aux données M/EEG. Nous avons montré que les essais conscients sont caractérisés par l'apparition d'un signal d'intention, représentant l'action requise, présent même lorsque l'on commet une erreur et qui constitue l'entrée des processus de détection d'erreurs. Ces résultats nous ont permis de proposer un modèle de détection d'erreur reposant sur la comparaison des deux flux d'information: le calcul non-conscient de la réponse motrice et la représentation consciente de la réponse requise.

In this thesis, we investigated the link between consciousness and metacognition. Metacognition, which can be defined as "cognition about cognition" constitutes the basis for evaluating, controlling and introspecting our one cognitive processes. While it is frequently assumed to be the hallmark of the conscious mind, this hypothesis should be empirically tested. Focusing on the simple, yet crucial metacognitive task of error detection, we studied how subjective conscious experience influenced behavior and magneto- and electro-encephalographic (M/EEG) brain response to errors. In a first study, we found that the ERN, a known brain marker of error detection was absent in subliminal conditions while subjects were still able to predict the accuracy of their decision slightly better than chance. These results suggest that two distinct performance monitoring processes co-exist: while all-or-none error detection, indexed by the ERN is present only in conscious conditions, confidence in one's response can still be computed under non-conscious conditions. To test whether these two processes were truly distinct, we replicated our study in a population of schizophrenic patients, who are known to present specific deficits in conscious conditions while their non-conscious processes remain unimpaired. Indeed, patients presented preserved metacognitive performance in subliminal conditions, while conscious error detection processes were altered, confirming that performance monitoring processes deployed consciously and non-consciously were computationally distinct. To further explore the difference between conscious and non-conscious error monitoring processes, we used the decoding methods of SVM linear classifiers applied on M/EEG. We showed that conscious trials distinguished from non-conscious trials by the emergence of a clear intention signal, representing the correct required action, present even when committing an error and influencing further error detection processes. These findings led us to propose an alternative model of error detection that relies on the comparison of two streams of information: non-conscious computation of the motor response and conscious computation of the required response.