Beyond introspective illusions, a brain computer interface approach to decision awareness

Benjamin Rebouillat

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'Ecole Normale Supérieure

Au delà des illusions introspectives, l'interface cerveau machine au service de l'étude de la conscience lors de la prise de décision

Soutenue par
Benjamin REBOUILLAT
Le 29 Novembre 2020

Ecole doctorale n° 158
Cerveau Cognition et Comportement (ED3C)

Spécialité
Sciences Cognitives

Composition du jury :

Valérian, CHAMBON CR HDR, Ecole Normale Supérieure	*Président*
Claire, SERGENT MCF HDR, Université de Paris	*Rapporteuse*
Elisa, FILEVICH CR HDR, Berstein Center for Computational Neuroscience	Rapporteuse
Lucie, CHARLES Post doctoral fellow, University College London	Examinatrice
Sid, KOUIDER DR HDR, Ecole Normale Supérieure	*Directeur de thèse*

1794
ENS
ÉCOLE NORMALE SUPÉRIEURE

BEYOND INTROSPECTIVE ILLUSIONS, A BRAIN COMPUTER INTERFACE APPROACH TO DECISION AWARENESS

BENJAMIN REBOUILLAT

PhD Thesis

Supervisor: Dr Sid Kouider

École Doctorale Cerveau Cognition Comportement

Département d'étude cognitive (DEC)
Laboratoire de Science Cognitive et Psycholinguistique (LSCP)

Équipe Cerveau et Conscience

September 2020

ABSTRACT

When we make a free choice, we feel conscious and in control of our decision processes. However, over the past decades, studies on introspection demonstrated that our self-knowledge faculties are crippled by illusory content. In Part i, we suggest that introspection can be framed as a hierarchically organized inference process and we proposed an innovative methodological approach to challenge this hypothesis. We used a free decision paradigm in which no high order nor low motor level processing were solicited. Further, we track in real time internal decision variables through a Brain Computer Interface (BCI), and probe both implicitly and explicitly participants' decision awareness. The present thesis investigates two main questions. First, what are the conditions for people to be aware of their impending decisions? Second, does people's introspections access genuine mental activity or are they pure retrospective illusions? Our results suggest that despite the general impression of a rich internal life, people are only partially aware of their impending decisions. If they can consciously track their upcoming decisions, they have no conscious access to those decisions' content. Yet, when recalling their recent choices, people can access internal representation of the chosen alternative. However, our results suggest that introspection has no privileged access to internal decision variables but rather stem from an integrative process involving both endogenous and exogenous cues. Introspective illusions thus reflect an imbalanced integration process, where weak and noisy internal variables are dominated by deceptive feedback. Overall, the present thesis provides new insights and methodological tools for the study of decision awareness emergence. Our results converge toward the idea that self-knowledge of decision is a hierarchically organized Bayesian inference process involving multiple cues.

Résumé

Nous concevons d'ordinaire nos choix comme conscients et sous notre contrôle. Toutefois, de nombreuses études montrent que nos processus introspectifs sont largement illusoires. Dans notre première partie, nous proposons que l'introspection peut être conceptualisée comme un processus d'inférence hiérarchique, et nous avançons une nouvelle approche pour en étudier les mécanismes sous-jacents. A cette fin, nous employons un protocole de prise de décision dans lequel les sujets ne peuvent accéder ni à leurs informations motrices, ni à des informations de haut niveau. En outre, nous mesurons les signaux neuronaux impliqués dans la prises de décision ainsi que la conscience que les sujets ont de leurs décisions. Cette thèse se penche sur deux questions: Premièrement sous quelles conditions peut-on être

conscient de ses décisions? Deuxièmement pouvons nous accéder à nos processus mentaux par l'introspection, ou cette dernière n'est elle qu'une illusion? Nos résultats suggèrent, qu'en dépit d'un sentiment de richesse subjective, nous n'avons qu'un accès partiel aux contenus de nos décisions. Si l'on peut savoir qu'une décision est imminente, son contenu échappe à la conscience. Toutefois, les sujets peuvent accéder à une représentation interne de leur choix a posteriori. Nos résultats soulignent cependant que cet accès reflète un processus intégratif au terme duquel notre introspection assimile à la fois des données internes et des informations exogènes. Les illusions introspectives sont dès lors le résultats d'une intégration déséquilibrée entre ces différents éléments. En conclusion, cette thèse offre de nouvelles perspectives ainsi que des outils méthodologiques pour l'étude de l'émergence de la conscience des décisions. Nos résultats convergent vers l'idée que la connaissance de soi est un processus d'inférence bayésien organisé hiérarchiquement et impliquant de multiples informations.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my PhD supervisor Sid Kouider. If I own Sid the very seeds of the present work, his guidance always preserved my independence in the conductance of my work. Though this period has been mark for Sid by the development of NextMind, I'm grateful he always managed to conciliate his new venture with our fundamental research, both in the interest he bears and the time he devolves to it. Aside from the regular and precious feedback he provided, I would like to especially thank him for his (very)patient guidance on improving my communication and writing skills. I would also thank him for having let me a great latitude in living my parallel life in Aikido. I hope our approach combining fundamental scientific questions and emerging technologies will continue to be developed for the enrichment of both side.

I would like to address special thanks to Jean-Maurice Leonetti, Nicolas Barascud and Robin Zérafa for their continuous assistance, guidance and advice in the handling of the brain computer interface. Without them, my life would have been so much more difficult that I'm not certain this work could have emerged in its present form. Also, I would like to express my thanks to Timothé Langlois Therin and Philip Sulewski that assisted during this thesis, I wish them good luck for their future academic journey. A special thanks to Matthieu Koroma, Camila Scaff and all those who shared my desk and more during those 5 years. Within the *Laboratoire de Sciences Cognitives et Psycholinguistique* (LSCP), I would like to thank Alejandrina Cristia for her support and advice in walking the delicate PhD road. I also want to thank Jérôme Sackur for his advice, notably when I was struggling with statistical issues. I also want to thank all his team that welcome us in their lab-meeting which have been a constant source of reflexions for me and help me step back and bring a fresh eye on my own work. My thanks also go to Elisabeth Pacherie, Valérian Chambon and their team who allow me to participate in their meeting. I found there dedicated people and a strong theoretical support for my own research. Finally, I want to thank Isabelle, Radhia, Vireack, Michel and Anne Caroline for their material and logistic support as well as for the informal daily chat.

Aside from academia, my thoughts go to the Aikido world. I would like to express my deepest gratitude to my sensei Okamoto Yoko Sensei and Fabrice Croizé Sensei as well as to all the members of Aikido Kyoto and Dojo des Guilands. I also want to mention and thank all those who participate daily to my guidance and practice, Christian Tissier Sensei, Hélène Doué Sensei, the Olympiades Aikido and the Cercle Tissier as well as the Central Island Aikido.

Finally, I want to thank my family and friends for their support during this period. My parents, my brother, my grandparents, Elena and all the Pontoisiens for their presence and their patience.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

BCI Brain Computer Interface

CB Choice Blindness

EEG Electro-encephalography

ERP Event related potential

FBCA Feature-Based Covert Attention

FRN Feedback related negativity

fMRI functional Magnetic Resonance Imaging

ITB Integration To Bound

MEG Magnetoencephalography

PDP Process dissociation procedure

PE Prediction-Error

RP Readiness Potential

SMA Supplementary Motor Area

ssVEP steady-state visual evoked potential

SWIFT semantic wavelet induced frequency tagging

Part I

INTRODUCTION

1

THEORETICAL INTRODUCTION

GENERAL OUTLINE

How conscious are we of our impending decisions? Such a question might seem trivial at first glance. Indeed, in our daily life, we usually think we are aware of our decisions and feel in control of their executions and consequences. However, that our introspective faculties are at least partially illusory is well established and documented since the early times of experimental psychology. In the present work, we investigate the conditions under which people can access the content of their decisions. In other words, we sought to understand whether people choosing between several alternatives are conscious of the option selected during the deliberation process but also when recalling their decisions. Another hypothesis could be that this self-knowledge is a pure narration retrospectively inferred to comply with other cues.

Indeed, decision content can be redundantly encoded in different brain signals. In one hand, internal decision process could inform people awareness about the choice their are about to make. Conversely, people might retrospectively infer their decisions based either on the motor preparation signals implemented to execute them, the external sensory consequences of their decisions or the memory of a long term planned decision.

To address this issue, we propose to study decision awareness through a Brain Computer Interface (BCI) based on the decoding of top-down covert attention. A first advantage of this approach is to remove any motor contribution from the decision process. Here, participants' decisions consisted in preferentially attending one over two overlapping stimuli at the center of the screen. Secondly, our BCI allows us to adapt the outcome of participants' decisions to their brain signals and thereby to control the external information relative to recent decisions. Finally, our setup allow us to measure through attention allocation a proxy for internal decision variables. We can thus confront this measure to participants' explicit reports of decision awareness, thereby assessing the condition under which internal decision variables are available to consciousness.

In Chapter 1 we review the theoretical and empirical literature on introspective illusions. We describe how the conception of illusion has evolved from a mere hindrance to experimental psychology to a natural outcome of introspection re-framed as an integrative process. We then discuss the implications for decision awareness of recent models and we propose that different aspect of the decision (i.e. the timing, the triggering and the content of the decision) are differently avail-

able at the different levels of the decision hierarchy (i.e. long term planning, concrete implementation or motor execution). We suggest that two mechanisms are responsible for the appearance of introspective illusions: First an imbalanced integrative process where internal decision variables are dominated by external cues. Second, a limited access to internal information at a certain level of the hierarchy during the decision process.

In Chapter 2, we show how BCI provides a novel approach to isolate a specific level of the decision hierarchy and study which aspects of decision are available to people's consciousness at each level. We review the cognitive mechanisms of attention upon which our BCI is built together with our approach to capture and exploit them. We then describe the implementation of real-time attention monitoring by our BCI allowing us to continuously track participants' decision processes. We further detail in Appendix A our visual stimulation which materializes the binary choice participants were asked to make. Our stimulation is designed to offer a simple picking decision to participants, thus removing any high level aspect from their decisions such as long term planning or preferences.

1.1 RETHINKING INTROSPECTIVE ILLUSION

What is the nature of our introspective illusions? When asked to explain our decision or justify our recent actions, we can flawlessly provide detailed justifications that motivated our behavior. Furthermore, we believe that these explanations reflect our privileged access to our mind. Stated in William James terms: "*All people unhesitatingly believe that they feel themselves thinking, and that they distinguish the mental state as an inward activity or passion, from all the objects with which it may cognitively deal.*" (James, 1890). Yet, a large part of the European philosophical tradition considers that our faculties for self-observation are partial if not illusory. Such illusion could be the substrate of a significant portion of our mental life, from slight misinterpretations of our common cognitive process (like finding yourself stating that "an idea came to my mind") to well established social facts (like believing that "I share my mind with a spirit" during possession ritual).

In Section 1.1, we review the ambivalent status occupied by introspective illusion in experimental psychology. Long considered as a mere noise contaminating experimental data, these illusions have been reconsidered in contemporary cognitive frameworks as an opportunity to study the mechanisms of introspection and consciousness. Furthermore, modern metacognitive paradigms, along with continuously refined methodological operationalizations of introspection, allow to consider illusion as an object of study *per se*.
In Section 1.2, we describe the integrative processes involved in the formation of introspection. Indeed, introspection over decision pro-

cesses integrate both endogenous and -potentially deceiving- exogenous cues. Such integrative accounts provide mechanistic explanations for illusions and transform the central question from *"is introspection accurate?"* to *"under which conditions are illusion occurring?"*. Finally, in Section 1.3, we suggest that various aspects of decisional processes are differently available to consciousness. If people might show reasonable faculty access to their intention to act (i.e when and whether they will execute their choices), they might remain unaware of the content of those impending decisions.

1.1.1 *A hindrance for early psychological science*

Contrary to a vision popularized by Watson, 1913, the advent of experimental psychology is a constant endeavor to diminish the source of illusion in introspective data (Costall, 2006). After a rapid overview of the epistemic and philosophical debates influencing early psychological science, we will show that early psychologists already had an advanced understanding of the phenomena of introspective illusions and that their practices of psychology were adapted accordingly. Consequently, the range of psychological investigation has been tightly linked to the methodological apparatus deployed to minimize the influence of illusions.

Introspective illusion: a epistemological obstacle for self knowledge

This is no recent idea that the "sense" by which people come to know themselves is deficient. Far from an exhaustive review of the philosophical debates on this subject, here are exposed some of the central ideas that influenced the birth of experimental psychology. A main dissociation in the methodological and theoretical approaches to introspection promptly emerged between the German and British tradition as a reflection of diverging conception of consciousness.

As expounded by William James (James, 1890) the very existence of a human predisposition for introspective observation has not in itself suffer much debate:

" *Introspective Observation is what we have to rely on first and foremost and always*. The word introspection need hardly be defined - it means, of course, the looking into our own minds and reporting what we there discover. *Every one agrees that we there discover states of consciousness*. So far as I know, the existence of such states has never been doubted by any critic, however skeptical in other respects he may have been."

However, whether this introspective observation is reliable at all in discovering internal mental states was, and remains nowadays a divisive issue. One one hand for the empiricist, reflection - the inner sense from which one obtains knowledge about her own mind- though indirect, is not subject to error (Locke, 1690). Empiricism will have a great influence on early experimental psychology and more particularly on the British tradition. As a result, British and Scottish schools

established psychological science as the study of the mind with introspection as its obvious and privileged method.

On the other hand German school was more reserved when it came to the use of introspective methods. This could be explained by a difference in the philosophical tradition that influenced German early psychologists. Following Leibniz's concept of *petites perceptions*, the mind processes an infinity of "small perceptions" and "integrates" them to give rise to an *aperception* (conscious perception) (Leibniz, 1765). Leibniz expresses here a fundamental dissociation between the sensory processing (perception) and its conscious experience (aperception) that will strongly influence the German school of psychology. On a similar ground, Kant considered introspection to be restricted to the *phenomenal* self, a representation of the mind that has no tangible existence outside of our thoughts and does not necessarily reflect the true nature of our mental life (Kant, 1910). Naturally, this philosophical background have promoted dissimilar practices of experimental psychology, emphasizing logic and rationality over introspective methods (Boring, 1953; Danziger, 1980).

Overall, contrary to the British perspective, the German tradition considered introspective illusion as a major hindrance for the foundation of experimental psychology. Attempting to address those concerns, Wilhelm Wundt delimited the use of introspection in psychological investigation.

Circumvent the illusion: early psychological approach of introspection

Wundt began by a fundamental distinction between Selbstbeobachtung (translated by "self observation") on one hand and innere Wuhrnehmung (internal perception) on the other. Selbstbeobachtung corresponds to the conscious observation of the self in the scientific sense (i.e. at the third person). It is thereby subject to -potentially deceiving- conscious inference. Conversely, innere Wuhrnehmung is compared with an immediate perception of internal mental life. Yet it can neither be reliable since it can become conscious and fall back in the first category. Therefore, Wundt recognized that the critics addressed to introspection are legitimate, but that they concern the Selbstbeobachtung, a reflexive process crippled with known illusory phenomena. Thereby, his efforts were directed toward preventing innere Wuhrnehmung to become conscious Selbstbeobachtung by rendering it similar to immediate sensory perception Wundt, 1888; Danziger, 1980.

Consequently, Wundt's methods aimed at minimizing the time interval between the internal perception and the report. Beyond avoiding memory distortion in introspective report, this shortening of perception-report time interval was meant to copy the observation in natural science which was considered as more direct and recognized as an

legitimate scientific method [1]. This led Wundt's to conduct his experiments preferentially on highly trained students. By training them, Wundt's hope was to turn students into "super introspecter" who could report their inner perception the same way they would have describe a visual scene (Wundt, 1883).
If this method reduces contamination by illusion and increases reproducibility of the reports, Wundt went a step further by drastically restricting both the extent of mechanisms affordable by introspection but also the very use of introspection (Danziger, 1980).

Introspective failures: impassable horizon of psychological science?

Indeed, Wundt considered that experiment could provide useful data only for a restricted portion of the psychological field. While perception and sensation were suitable for introspection, social psychology or voluntary decision were subject to an insurmountable degree of illusory processes and should be addressed by historical, anthropological or sociological approaches. Thus, "introspection" in Wundt's laboratory was mostly reduced to simple judgments like stimulus intensity, size or time perception while more complex reports like qualitative description were almost banished. Furthermore, contrary to a popular belief, Wundt primarily relied on objective methods like behavioral measure and language analysis (Costall, 2006; Danziger, 1980).

Overall, until the early 20th century the introspective illusion has been conceptualized as a hindrance for psychological investigation. The strong constraint imposed by early psychologist to avoid pollution of their measures by illusion phenomena have drastically curtailed the reach of their research and the attempts to overcome this problem ended in a drastic discrediting of introspection and the advent of behaviorism (Blumenthal, 2001). However, contrary to the simplistic view that behaviorism replaced "introspectionism" in the early XXth century, introspection has never been a dominant practice in the psychological field, neither will it disappear with the expansion of behaviorist school (Costall, 2006). Instead, introspective methods will be progressively refined to settle on more rigorous ground. Furthermore, a regain of interest of consciousness in the second half of the XXth century will accompany a progressive shift in the conceptualization of introspection. From a research method, introspection will become an object of interest *per-se* and the associated illusion will be used as an observable manifestation of its mechanisms.

1.1.2 *An object of study per se*

Although the use of introspection did not completely disappear, its use has been more discrete in a large first half of the XXth century

1 We can find in James Sully's *Illusions of introspection* the idea that introspective illusion can affect internal perception and thus the psychological data the same way perceptual illusion are affecting external observation (Sully, 1881)

(Costall, 2006). The 60's however, are marked by a resurgence of interest for consciousness and the re-emergence of studies relying upon introspective reports. This resurgence of interest for subjective variables has been permitted by both methodological and theoretical re-conceptualization of introspection that, above being a suitable method, will become an object of study *per-se*. These early work will be followed by a characterization of metacognition first in child development and soon extended to all fields of psychology, formalizing a framework to study introspective illusion.

Methodological and theoretical rehabilitation of introspection

The emergence of modern techniques in neuroscience has deep implications for the use of introspection in experimental psychology. Indeed, regularities between measurable brain features and verbal reports can now be systematically assessed (Gazzaniga and Sperry, 1967). Yet, introspection has not only taken advantage of new neuroimaging approaches but also benefited of a continuous methodological refinement permitting its systematization while preserving its subjective dimensions (Sackur, 2009).

The re-conceptualization of introspection in the 60's is nicely illustrated by the seminal study of Sperling on brief visual presentation. Considered as one of the founding paper of the cognitive science, Sperling, 1960 associated careful timing of the visual stimulation with a dual condition verbal report demand. In a nutshell, after a brief visual exposition to a grid of letters, participants claim to perceive all the letters but are only able to report 4. Yet, when asked to report only one line of the grid (designated *after* the visual presentation), participants recover the ability to report the 4 letters of this specific line. Although the consequences of this work are still debated (Kouider et al., 2010; Block, 2011), it unravels a distinction between conscious experience and the information available to report.

Thus, introspection progressively acquired a dual status, both as an object of study and as an investigation tool. Associating these two aspects opened new questions (Nisbett and Wilson, 1977): (a) What is the basis of [...] accurate reports? (b) Are accurate reports fundamentally different in kind from inaccurate ones? (c) Is it possible to specify what sorts of reports will be accurate and what sorts will be inaccurate? Although those questions receive contrasted answers (Ericson and Simon, 1980; Nisbett and Wilson, 1977), they paved the way to the formalization of monitoring and control processes over mental states under the concept of metacognition.

Metacognition: a reinterpretation and emancipation of introspection

Metacognition has been defined as a combination of introspective and control processes. The first definition will be given by Flavell, 1976 as follows:

"The knowledge of each individual regarding its own cognitive processes and products [...] Metacognition refers to active control and subsequent regulation of the objects of knowledge (mental contents)".

From this definition, the metacognitive processes associated with a wide range of cognitive process will be described. Starting with learning in children (Flavell, 1979), metacognition will be reported for attention(Yussen and Bird, 1979), memory (Nelson and Narens, 1990), decision making (Yeung and Summerfield, 2012) or social cognition (Frith, 2012) among many other cognitive processes. A second structuring account of metacognition is proposed by Nelson and Narens, 1990 which framework relies on three principles:

- *The cognitive processes are split into two or more specifically interrelated levels.*
- *The meta-level contains a dynamic model (e.g., a mental simulation) of the object-level*
- *There are two dominance relations, called "control" and "monitoring" which are defined in terms of the direction of the flow of information between the meta-level and the object-level.*

The object-level entails here the mental process involved in cognitive task such as memory, learning or decision making. On the other side, a meta-level that monitor and control the object-level processes. Furthermore, Nelson and Narens, 1990 proposed that meta-level implement control a representation (or model) of the object-level (Conant and Ross Ashby, 1970).Thus the formalization of metacognition render introspection indissociable from control processes. Thereby, introspection could not only be illusory but also directly impact the system through control instances.

Interim conclusion

Overall, introspection illusions were a documented phenomena since the foundation of experimental psychology. The debate about their suitability as a tool for experimental psychology have progressively evolved in interrogation about their underlying mechanisms. Indeed, methodological refinements together with the introduction of the concept of metacognition in psychology allowed us to consider illusion as a rightful object of investigation. Yet, the interrogations about the nature of introspection remain centered on its accuracy. In the next section, we will review how integrative account of metacognition of decision conciliates empirical findings in support of both illusory and accurate introspection. This line of research further redefines the relevant questions on illusion, centering them on the conditions under which an information impacts the awareness.

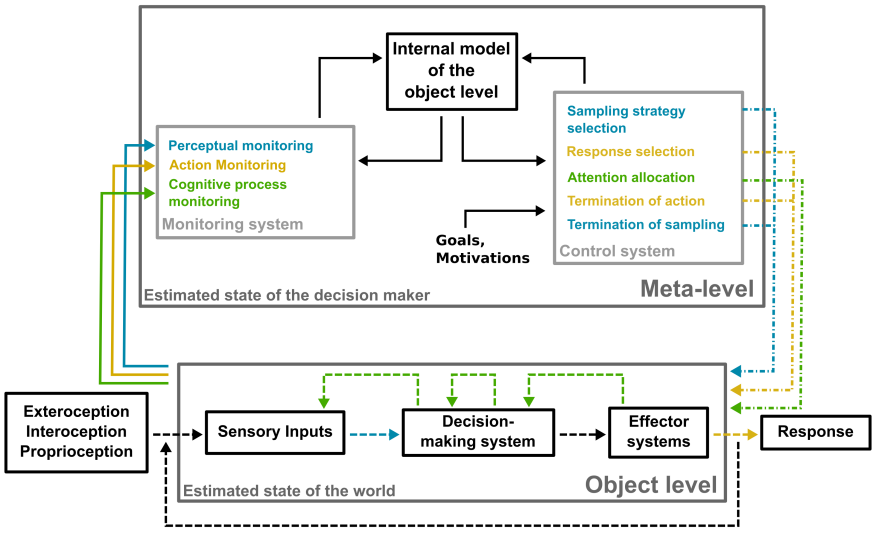

Figure 1: Metacognition principles applied to decision process. *Figure is adapted from Nelson and Narens, 1990; Qiu et al., 2018; Heyes et al., 2020.*
The object level computes first order representation of the world. Dashed line show (unconscious) information flow within the perception-decision-action loop. Perceptual information is shown in blue. Motor signals are shown in yellow. Unconscious cognitive process at the object level are shown in green.
The meta level computes second order representation such as the decision maker state. It encapsulates an internal model of the object level that is constantly updated through the ascending monitoring system (solid lines). Furthermore, the meta level supervised the object level through a descending control system (half dashed lines). Text color in the meta level corresponds to the associated states and processes supervised at the object level.

1.2 FROM ILLUSION TO INTEGRATIVE PROCESS: BEYOND THE INTROSPECTION ACCURACY DEBATE

1.2.1 *Illusion: the default mode of self-knowledge?*

In the metacognitive framework, introspection is described as a central mechanism impacting a wide variety of cognitive processes through its associated control instance. Imperfection of introspective processes would result in under-optimal control of the behavior as participants fail to monitor their precise mental states. However, a rapidly growing literature formulated a substantially different hypothesis. Illusion according to them, could be the default mode of introspective processes. According to such an account, introspective illusions would not be an exception in a normal, accurate monitoring process but rather the norm of our mental life. One major consequence of such an account is that our impression of consciously controlling and causing our behavior would be illusory.

Illusion is the common output of introspection.

This conclusion has been proposed by Nisbett and Wilson, 1977. In their seminal study, they concluded after an extensive review of the social psychology literature that if people can access their mental states (e.g. "I am cold", "I am hungry", etc.), their reports on cognitive processes are confabulated. As an example, participants taking a placebo pill could endure more electric shocks but attribute this enhanced resistance to their personal history or unrelated event rather than to the pill (Nisbett and Schachter, 1966). Such experiments illustrate that people remain usually unaware of the external stimuli affecting their behavior. Therefore, even accurate introspection would merely reflect the accidental correspondence between the cognitive actual process and the unrelated personal narration that people built from their experiences and implicit causal theory Nisbett and Wilson, 1977; Wilson and Dunn, 2004.

A major critic addressed to this hypothesis is that it relies on a dubious distinction between cognitive process and cognitive states Ericson and Simon, 1980; Reyes, 2015. In effect, Nisbett and Wilson, 1977 describe cognitive process essentially as the "cause" (Neisser and Becklen, 1975; Miller, 1962) that guide a behavior or that link external event with participants actions. However, causality is largely criticized as being a theoretical intractable construct. It is therefore unsurprising that this concept is not a consciously accessible substrate of participants' mental life (Engelbert and Carruthers, 2010).
Yet, in spite of the challenges addressed to Nisbett and Wilson, 1977, recent innovative protocols provide additional cues that people barely know their own thinking and decision (Johansson et al., 2005).

In Choice Blindness (CB) paradigm (Johansson et al., 2005), participants have to choose between two face pictures the one they prefer. Immediately after, they are given the chosen pictures and asked to provide justifications for their decision. Crucially, a legerdemain was used on certain trials to present instead the non preferred picture. In a large majority of manipulated trials, participants continue to provide justifications that could even be in clear contradiction with their original choice (e.g. "I choose this face because I prefer blond women" while the original choice designated a brown-haired woman). Even during funneled debriefing, participants presented no awareness about the manipulation.

Early attempts to unravel cognitive mechanisms supporting CB episodes compared the introspective reports following manipulated and non-manipulated conditions. Interestingly, linguistic analysis including psychological rating of reports, word-frequency analysis and latent semantic analysis did not succeed in differentiating between reports across conditions (Johansson et al., 2005; Johansson et al., 2006; Johansson, Hall, and Sikström, 2008). Such absence of result may reflect the similarity in the cognitive processes involved during reports

following manipulated and normal feedback. Thereby, those analyses cast doubt on the propensity of all introspective reports to actually reflect participants' actual motivations [2] rather than retrospective illusions.

Thus, illusion could constitute people's "normal psychological life" during introspection. Yet, if introspection does not reflect the actual cognitive processes underlying people's action, then the question of the basis of its illusory content should be asked.

Illusion stems from retrospective interpretation.

In line with Nisbett and Wilson, 1977, Wegner, 2002 argues that people have poor knowledge about the cause of their actions. Moreover, their conscious experiences of free will would result from a retrospective narration based on elements such as the perception of the executed movement or congruent instructions to act. This hypothesis is exemplified in the seminal Libet et al., 1983 study where participants are asked to perform "when they feel an urge to move" a button press. Libet compared the timing of the appearance of cerebral activity preceding the voluntary action (namely the Readiness Potential -RP-) with the appearance of the conscious experience of wanting to perform this action (W time). Surprisingly, W lag several hundreds of milliseconds behind the initiation of RP leading Wegner and others to suggest that introspection about action is mostly a retrospective process.

This account is further developed by Wegner's own empirical research on apparent mental causation. In a series of studies, he shows that people can feel in control for an action when they obviously have no part in it. For example when subjects are instructed to move their hand while observing an experimenter's hand moving, they will report a higher feeling of control on the movement than without instructions[3] (Wegner, Sparrow, and Winerman, 2004).

Following Wegner hypothesis and epiphenomenalist account (Huxley, 1874), both action and thought would have unconscious origins that independently drive them. Then, a thought to be retrospectively considered as the conscious causal origin of an action must meet three criteria (Wegner, 2002; Wegner, 2003):

- *Principle of priority:* The thought has to appear in consciousness before the action.
- *Principle of consistency*: The thought has to be consistent with the unfolding action.

2 In Johansson et al., 2005 own words: "Confabulation could be seen to be the norm and truthful reporting something that needs to be argued for"

3 Similarly, in a Ouija-like setup, while the experimenter induce all the movement of the pointer, participants can feel in control of the pointer notably if the movement match participant's instruction (e.g. to point a specific item on a picture Wegner and Wheatley, 1999; Wegner, 2004)

- *Principle of exclusivity*: The thought has to be the most likely origin of the action.

Yet, both Libet and Wegner empirical and theoretical findings have been largely criticized. Among theoretical, methodological and even statistical criticisms (Dennett, 1991; Walter, 2014), two axes of discussion are of interest for the present work. First the experimental setup they used were often meaningless. Both the setup and the instructions had little if any resemblance with real life decision scenarios. This is notably Libet's experiment (Libet et al., 1983) where the instruction of waiting an "urge to move" has been widely criticized. Furthermore, the choices made by participants were not involving the distant planning components that constitute every day decisions (Mele, 2014). Such over simplification of the decision-making protocols are thought to shortcut certain psychological mechanisms (Robinson, 2019). A second axis of criticism concedes the existence of situations in which introspection might be a retrospective illusion while the behavior is mainly driven by unconscious processes. However the existence of such episodes do not imply that this situation prevails and even less that it constitutes the standard for our self knowledge faculties.

1.2.2 *Beyond illusion: integrative perspective of introspection*

To demonstrate that introspection is not a pure retrospective illusion, one approach is to distinguish the cues participating to the introspective process given their origin. On one hand, introspection can be based on exogenous cues like perceptual or contextual information. Choice blindness illustrates such a situation nicely as participants' introspection mostly reflects the presented outcome rather than their preferences. On the other hand, introspection could rely on endogenous cues such as people preferences, prior belief or motor preparation signal. Under this framework, introspection can still be view as a monitoring process, but its accuracy will depend on the cues it encompasses. Importantly, the origin of a cue does not guarantee its validity as confabulation can be based on people's preferences or prior beliefs (Nisbett and Wilson, 1977; Johansson et al., 2005). Yet, such distinction allows us to consider introspection as an integrative process where both endogenous and exogenous cues can be involved.

As a matter of fact, endogenous information has been shown to impact participants' decision awareness. Indeed, motor preparation signals influence the awareness participants have of their intention to perform an action. Recent empirical and theoretical studies proposed an integrative model of certain introspective features such as the sense of control over our own actions.

Introspective process involve endogenous information

Inspired by engineering closed loop models, the motor prediction view (Frith, Blakemore, and Wolpert, 2000) holds that control pro-

cesses of motor action are based on the coupling of inverse and forward models through a series of comparators. From a desired state, an inverse model computes the motor command that would lead to this state. The motor command is then sent to the effector motor system but also to a forward model that will emulate the predicted state of the system after the command execution. Three comparators serve as control instances (see Figure 2):

- Predicted/Desired states: Finds the optimal command to attain system desired states.
- Predicted/Actual states: Filters in the sensory feedback what is attributable to the system action and what is caused by other, external sources.
- Desired/Actual states: Detects mismatches and improves the forward model to reduce them.

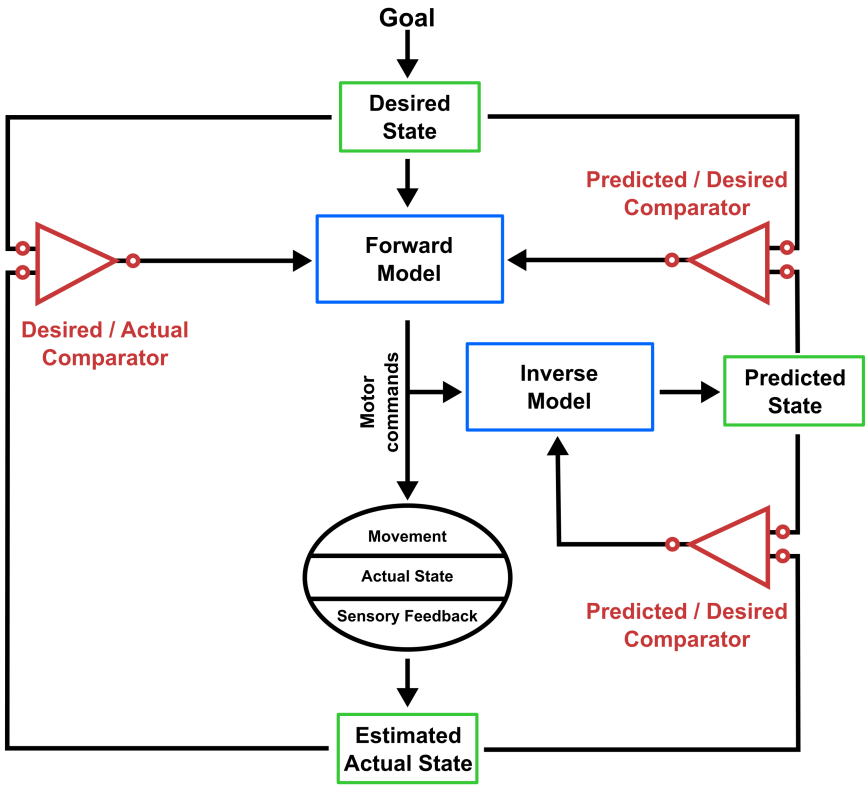

Figure 2: The motor prediction framework. *Figure is adapted from Frith, Blakemore, and Wolpert, 2000.*
The model is described in Section 1.2.2. Internal models are shown in blue. Comparators are shown in red. The different representational states involved in the control of the action are shown in green.

One of the main idea of the motor prediction view is that the output of these comparators are the principal cues for the sense of agency (the feeling to be at the origin, in control and responsible for the consequences of our action). The awareness of initiating one action would

come from a comparison between predicted and desired states rather than a retrospective interpretation of the current states. In line with this hypothesis, the sense of agency is gradually reduced as discrepancies between predictions and sensory re-afferences increase (see for example Fourneret and Jeannerod, 1998; Sato and Yasuda, 2005). Furthermore, people are aware of their intentions to act at relatively early stage of motor preparation. This awareness is specifically associated with the RP (Parés-Pujolràs et al., 2019) and is impaired by posterior parietal lesions (Desmurget and Sirigu, 2009).

Overall, those findings temper the illusionist account of introspection of our decisions. However, the motor prediction view and the cognitive reconstruction account are not mutually exclusive. Indeed both endogenous and exogenous cues could interact to form conscious experience of the diverse aspects of decision, from deliberation to their execution.

Wrapping up endogenous and exogenous influences

Can endogenous and exogenous cues simultaneously contribute to introspective processes? Research on awareness of decision and action have largely operationalized this question with intentional binding measure. Intentional binding refers to the subjective temporal attraction between a voluntary action and its sensory consequences. This measure offers an indirect access to awareness of action and the sense of agency (Haggard, Clark, and Kalogeras, 2002) and has been shown to correlate with introspective reports of agency (Imaizumi and Tanno, 2019). As such, it could be considered as an operationalization of introspection. Interestingly, intentional binding is not only modulated by predictive motor process but also by retrospective inference and outcome probability (Moore and Haggard, 2008; Moore, Wegner, and Haggard, 2009; Evans et al., 2015). Crucially, the contribution of endogenous and exogenous cues is modulated by their respective reliability (Moore, Wegner, and Haggard, 2009; Wolpe et al., 2013). As an example, motor signals contribution will be stronger (i.e. drive a larger subjective temporal compression between action and outcome) if the outcome is unpredictable (Moore and Haggard, 2008).

Following these empirical findings, theoretical integrative accounts of the sense of agency have been proposed (Moore and Fletcher, 2012; Synofzik, Vosgerau, and Newen, 2008; Legaspi and Toyoizumi, 2019). In line with multi-sensory perception models (Ernst and Banks, 2002; Knill and Pouget, 2004), recent models propose that sense of agency results from a Bayesian integration. According to this view, the reliability of motor preparation and exogenous cues together with the prior expectation of the participants will mediate their respective influences on the the sense of agency (Moore and Fletcher, 2012).

Bayesian approach has crucial consequences for the conceptualization of metacognition or introspection. First, it considers internal vari-

ables as a noisy source of information for metacognitive processes. Thereby it rejects the idea of a privileged access to one's own mental life. Second, since external cues also contribute to the formation of introspection, the central question becomes *whether and how the different cues contribute to metacognitive processes*. One framework to address this question can be the predictive coding, which proposes that the processing of the different cues is distributed in a Bayesian hierarchical organization (Srinivasan, Laughlin, and Dubs, 1982; Mumford, 1992; Rao and Ballard, 1999; Friston, 2008; Friston et al., 2017).

The predictive coding framework

What is perception? Predictive coding rely on the core concept that the brain is constantly generating and updating a model of the environment (Friston, 2010). More precisely, the brain encodes the statistical regularities of the environment in a hierarchy of top-down generative models of various temporal and space scales. Those models *predict* and *suppress* inputs ascending from lower levels. Predictions (prior) are compared to the lower level inputs (likelihood) and if their discrepancy (i.e. the prediction error PE) is sufficiently large, it causes the generative model to be revised. Otherwise, a negligible PE reflects an accurate model, which posterior probability is thereby increased. Thus, perception under predictive coding is mostly a prediction process aiming at reducing PE[4], where sensory inputs only affect percept in the form of PE propagating up the hierarchy (Hohwy, 2013; Friston, 2012; Wikipedia, 2020).

What are attention and metacognition? When adjusting a model, the influence of top down prediction[5] and bottom-up information[6] are weighted by their relative precisions. Precision can be understood as the confidence attributed by the hierarchical model to the prior belief and sensory evidence respectively. Therefore, the brain does not only predict the content of the sensory information but also the precision -or confidence- that contextualize this information at each level of the hierarchy. The parameters encoding for precision are called "hyperparameters" and the prior beliefs about them "hyperprior" (Friston, 2008; Friston, Lawson, and Frith, 2013). Crucially in the framework, a key metacognitive feature is to optimize hyperpriors or, in other words, to optimally infer the relative confidence attributed to sensory inputs and prior beliefs (Asai, 2017). From a psychological point of view, precision modulation is associated with attentional gain. Predicting that a sensory input is precise is equivalent to attributing attention to its processing (Friston, 2012; Picard and Friston, 2014). Indeed, if a source is precise, we attribute attention to its information and adapt

4 In the predictive coding theory, PE reduction is proposed as the only thing the brain ever does Hohwy, 2013. Doing so, the brain reduces its free energy and maintain itself in a restricted number of states, which entail beliefs about the world. Restricting the potential number of occupied states is suggested as a solution for living being to stay in a non-equilibrium steady state (Friston, 2010)

5 The internal prior beliefs

6 Information traveling from the environment up to the cortical hierarchy

our prediction according to them.

What is a decision? Instead of updating an internal model, a second option to reduce PE in the brain is to act on the environment. One can decide to move by formulating a high level prior about his body (e.g. "I am moving my arm"). This prior will generate predictions about the current state of the world and the expected sensory consequences of the action. In turn, this prediction will drive the behavior until the movement is performed (Clark, 2015). But since the arm is not moving yet, the PE resulting from the discrepancy between sensory inputs and prediction is high. To prevent PE from updating the prior and thereby cancel the movement, active inference theory proposes that the brain minimizes PE impact by attenuating the precision (i.e. withdraw attention) of sensory inputs until the action terminates. Thus, our own decisions are characterized by the reduced attention we attribute to their sensory consequence (Asai, 2017; Picard and Friston, 2014; Vasser et al., 2019). Again here, a metacognitive feature (the monitoring of one's decision process) is described as an optimization of precision expectation.

What is an illusion? Metacognitive illusions can be explained in terms of deficient allocation of precision (Picard and Friston, 2014). Reducing the precision of sensory input compared to prior belief can simulate illusion of control over others' movement. Indeed such reduction could label other's movement as "attenuated" and therefore controlled. On the contrary if one failed to attenuate the precision of a self-generated movement, she will experience alien control of her limb (Picard and Friston, 2014; Brown et al., 2013; Asai, 2017). Propagation of these non-attenuated PE can lead to actualization of high order models to account for them, resulting in delusional belief as observed in schizophrenia (Fletcher and Frith, 2009).

Interim conclusion

As we saw, recent models of metacognition allow us to revisit the question of introspection accuracy. Rather than a read out of internal variables eventually contaminated by external elements, introspective processes are described as probabilistic inferences about properties (i.e. the precision) of the mental states (i.e. prediction on sensory inputs) (Friston, Lawson, and Frith, 2013). In this context, metacognitive failures can be re-interpreted as failure to estimate the respective influence (or precision) of the different sources of information contributing to a mental state. [7].

In the next section, we propose a theoretical and empirical framework to investigate whether people are aware of the content of their

7 In line with this theory, participants reporting their confidence are thought to monitor precision (Meyniel, Sigman, and Mainen, 2015). Indeed, it has been proposed that precision in Bayesian neural computation and the confidence in psychology are related like a set data and its summary statistic (mean, standard deviation, median etc...)

decisions. In line with study on intention to act (Schultze-Kraft et al., 2020), we suggest to confront participants' introspective reports to their internal decision evidence. Furthermore, we propose that the hierarchical organization of the processing of exogenous and endogenous information impacts the quality of introspection.

1.3 MONITORING AND CONTROLLING DECISIONS

While a decision process is unfolding, several features have to be computed. Brass and Haggard, 2008 distinguish three main aspects forming a decision to act, namely the when (timing of the action), whether (triggering of the action) and what (content of the action) aspects. In this section we will see that those aspects are thought to be hierarchically processed and we will propose that this hierarchical organization impact the availability of the different features to awareness. Recent studies demonstrated that people can be aware and in control of the initiation of their decision execution (when and whether aspects) even for arbitrary choice essentially processed at lower level (Schultze-Kraft et al., 2016; Schultze-Kraft et al., 2020; Parés-Pujolràs et al., 2019). Here, we will suggest that awareness and control of the what aspect of a decision is dependent on higher order levels of processing. In other terms, people could be conscious and in control of the object of their choices at the level of decision planning but not during concrete implementation of the decision execution.

1.3.1 *Reframing decision*

1.3.1.1 *Theoretical account*

Voluntary decisions along with intention to act have recently been suggested to be hierarchically organized processes (Mele and William, 1992; Pacherie, 2008; Brass, Furstenberg, and Mele, 2019). These theories distinguish between distal intentions that are high-level, motor intentions that concern low level motor execution and proximal intentions which implement motor intention according to distal planning.

At a higher level of the hierarchy, the distal intention encompasses the intention to act in the future according to long term goals (e.g. wanting to eat fish). Distal intentions are conceptual and keep a degree of abstraction allowing to adapt concrete decisions to the situation. Furthermore distal intentions play a metacognitive role toward lower levels of the decision hierarchy since they monitor and control the decision process downstream (the series of decisions leading to eat fish). Therefore distal intentions have to be maintained in memory and are thought to be available to conscious access.

At a lower level, proximal decisions imply the more immediate future and serve the achievement of higher level decisions (e.g. go fishing on the river). Proximal decision thus integrates higher order intention and the present context to adapt decision to the environment.

In addition, proximal decisions are in charge of controlling the execution of decisions by monitoring motor implementation. This type of intention is thought to have a brief lifespan as they are bound to the impending action. Thus, as far as the proximal level is responsible for "initiating, sustaining, and guiding intentional actions" people may have only limited conscious access to the content of their proximal decisions (Mele and William, 2009).

After reviewing recent theories on the neural implementation of decision, we will suggest that people could only be partially aware of their proximal decisions content (Kouider et al., 2010). Indeed, while people could be aware of whether and when they are about to act at the proximal level, awareness about the content of the decision would depend on a higher level of the hierarchy.

1.3.1.2 *Neural implementation*

Popular models of decisions based on perceptual evidence rely on integration-to-bound (ITB) principles by which the perceptual evidence is sampled and accumulated until reaching a threshold, thereby triggering the decision (Bogacz et al., 2006; Mulder, Van Maanen, and Forstmann, 2014). Diverse implementations of ITB include race model (Usher and McClelland, 2001), where the evidence is separately accumulated for each choice alternatives, or drift diffusion model (Ratcliff and McKoon, 2008), where the accumulator represents the relative evidence of one decision over the other. Those models accurately predict choices but also other variables like reaction time in simple 2 alternatives forced choices paradigms (Ratcliff et al., 2016).

Recent studies suggest that internal based decisions also rely on ITB mechanisms. Yet, instead of sampling and accumulating perceptual evidence, the decision to act is reached by accumulating the neural random fluctuations of the motor system (Schurger, Sitt, and Dehaene, 2012; Murakami et al., 2014). Furthermore, these internal evidence and perceptual evidence based decisions exhibit a common encoding in the fronto-parietal cortex (Wisniewski, Goschke, and Haynes, 2016). Indeed, similar to how external evidence accumulation drives the perceptual-based choices (Soon et al., 2008; Bode et al., 2011), internal random fluctuation of the decision neural precursors are accumulated to break the symmetry between the alternative options and trigger endogenous decision (Schurger, Sitt, and Dehaene, 2012; Maoz et al., 2013; Murakami et al., 2014; Furstenberg et al., 2015; Deco and Romo, 2008).

In the following paragraph, we will see how those approaches could help us to interpret recent findings linking neural decision precursors and participants' awareness.

1.3.2 *Introspect decision timing and execution*

Already evoked in the present manuscript, the readiness potential (RP) is a ramping EEG activity over pre-SMA (a frontal motor areas) initiated around 1 s before voluntary movement onset (Haggard, 2008). Early interpretations suggested that RP would reflect unconscious initiation of -not so- "voluntary" movement. Yet, considering that endogenous decisions rely on accumulation of neural random fluctuation, recent study showed that RP-like signals can be obtained by averaging the leaky accumulator across trials (Schurger, Sitt, and Dehaene, 2012). Furthermore, active reduction of random fluctuation has been proposed as a key neural signature of voluntary decision process, independent from motor execution (Khalighinejad et al., 2018), which could result from a controlled reduction noise sampled by the accumulator during voluntary decision preparation (Khalighinejad et al., 2018; Khalighinejad et al., 2019). Having identified these precursors of voluntary action, the question becomes whether and how the information they convey can be accessed and exploited by participants.

Since Libet et al., 1983, the faculty to veto is regarded as a hallmark of conscious access to decision processes (Block, 2007). Even after initiation of a RP, participants are still able to veto their decisions and refrain from acting (Libet et al., 1983; Schultze-Kraft et al., 2016). Externally triggered cancellation have been found to be possible up to 200 milliseconds before the movement onset (Schultze-Kraft et al., 2016). Recent study ask people to perform a self-paced button press and regularly probe them to report their intention to move. They found stronger RP-like activity in the period preceding an intention to move (Parés-Pujolràs et al., 2019). Furthermore, innovative approach uses Brain computer Interface to trigger probes when detecting a RP (Schultze-Kraft et al., 2020). Those probes inform the participants to either move or refrain from moving. Participants where then asked to report whether they were intending to move at the probe appearance. This study shows that both go-probe and the presence of RP increase the probability of reporting an intention to move. Together, these studies demonstrate that participants have at least some access to the neural precursor of their voluntary action, including the random neural fluctuations thought to trigger action (Parés-Pujolràs et al., 2019). Especially, people seem to access and thereby control the initiation of their voluntary actions.

Brass, Furstenberg, and Mele, 2019 recently proposed a model of decision wrapping together hierarchical organization, accumulator model and empirical evidence on conscious access. The model suggests that high level instances of the decision hierarchy implement the accumulator parameters, and that threshold crossing triggers the proximal decision (Kang et al., 2017; Schurger, 2018) to act rather than the action. In this context, distal decision is thought to determine the following parameters:

- *The accumulated variable:* The model proposes that the only difference between externally and internally driven decisions is the nature of the accumulated variable. Externally driven decisions would rely on sampling of perceptual evidence while sampling of internal random fluctuations supports endogenous decisions.
- *The threshold(s) :* Setting the threshold controls the timing of the future decision. In an internally driven decision it ensures that the decision can be achieved within a given window.
- *The initial bias:* Bias represent initial preference for a given alternatives which might reflect prior beliefs (e.g. preferences, choice history ...) or perceptual primes. Importantly in the present model, arbitrary picking choice and meaningful choice can be placed on a continuum depending on how the threshold is crossed. A purely arbitrary choice results from random noise accumulation driven threshold crossing while the bias plays a dominant role for triggering meaningful decisions.

By accounting for the emergence of proximal decision rather than decision execution, this "conditional intention and integration to bound model" (Brass, Furstenberg, and Mele, 2019) endorses a large amount of theoretical and empirical findings. In particular, it leaves room for a time between the emergence of decision in consciousness and its execution, thus providing a mechanistic explanation for vetoing opportunities, even for arbitrary decisions. From an evolutionary perspective, the faculty to veto an impending decision might always be advantageous even for arbitrary choice as it permits environmental adaptation and danger avoidance. It is therefore unsurprising to find an access to neural precursors of the intention to act even at a relatively low level of the decision hierarchy. However, when the choice alternatives are equivalent, consciously controlling the content of the proximal decision might be irrelevant.

1.3.3 *Introspecting the decision content*

Neural signals predicting the content of motor and abstract decisions (e.g. mathematical operation) have been identified using fMRI and Libet-like paradigm (Soon et al., 2008; Soon et al., 2013). Of course, when people have to make meaningful decisions reflecting their preferences, they should be aware of their internal deliberations and in control of their decisions. One way to study the awareness of the decision content is to use choice blindness (CB) paradigm with different choice alternatives. A simple operationalization of people awareness consists in counting the number of manipulated trials going undetected by participants (CB episodes). In line with our hypothesis, decisions involving familiar choices (e.g., known brands, political preferences, etc) are rarely followed by CB episodes (Hall, Johansson, and Strandberg, 2012; Sauerland et al., 2014; Somerville and McGowan, 2016; Rieznik et al., 2017). Moreover, expert guidance during a decision prevents occurrence of CB episode (Petitmengin et al., 2013).

Noteworthy however, rewarding a decision might not necessarily improve participants' conscious access (Hall et al., 2010). Yet, this absence of effect might also be attributable to social factors (Reyes et al., 2018, *unpublished data*).

To investigate whether this conscious access remains when no distal component guide the decision, a strategy consists in using arbitrary picking choices. Picking is differentiated from choosing as it involves no reasoning (Ullmann-Margalit and Morgenbesser, 1977). Picking are conceptualized as reflecting purely proximal decision, free of long term intention (Pacherie, 2008; Mele and William, 2009; Furstenberg et al., 2015). In a sense, a vast panel of decision paradigms used in experimental psychology can be classified under this label, often provoking critics on their lack of realism compared to real-life choices. Yet, as we emphasized in the previous section, there are good reasons to believe that picking and choosing reflect two extremes of the same evidence accumulation processes. In line with this hypothesis, changes of intention after subliminal priming have been shown during picking decisions. Despite being primed to perform a (e.g. left) button press, participants sometimes "change their mind" and push the right button (Furstenberg et al., 2015). This phenomenon could be explained by positing that, despite the bias imposed by the prime, accumulation of random neural fluctuations dominate the process of threshold crossing, resulting in the opposite-to-bias decision to emerge.

In the present work, we further isolate proximal decisions by removing lower levels of hierarchy in a picking scenario. Indeed, our participants did not perform any motor action which they could monitor to gain information about the content of their impending decisions. Instead, they were asked to decide to preferentially attend one of two overlapping items on the screen. Under the predictive coding theory, to attend one item is equivalent to increase the expected precision of the PE resulting from the perception of this item. Thus, participants' 'choices can be reformulated as attributing higher precision to one of the two alternative items. Since the precision are equal among choices, the attention allocation must result from endogenous factors (Clark, 2017; Ransom, Fazelpour, and Mole, 2017). Those endogenous factors usually entail desire and motivations but are restricted in a picking scenario to proximal aspects of the intention. To summarize, we isolate the proximal level of decision by removing on one hand the distal aspect of decision using a picking scenario, and the sensorimotor cues on the other hand using an attention allocation driven choice.

In Chapter 2, we will see how we track participants' feature based covert attention allocation using decoding methods. This highly non ecological paradigm allows us to investigate whether and how participants could become aware of their impending decisions at the proximal level, i.e the level of concrete implementation of decision.

2

CAPTURING DECISION SIGNALS IN REAL TIME.

As we have seen in Chapter 1, research on the conscious experience of volition addresses a intriguing contradiction. On one side, a growing corpus of studies using imaging methods have identified brain signals predicting the content of a forthcoming decision (see Section 1.3). However, cognitive psychological studies on introspective abilities have shown that people are differentially aware of the different aspects of their impending decision. While they can prospectively access and control their intentions to act, they are remarkably unaware of their impending decision (see Section 1.2). This apparent contradiction has usually been explained in terms of the illusory introspective process underlying the conscious experiment of our intention. Nevertheless, recent innovative studies unravel genuine metacognitive access to internal motor preparatory signals as the action unfolds. Yet when retrospectively monitoring their intentions, people's introspective content would be dominated by the decision they took and the perceptual feedback they received (Schultze-Kraft et al., 2020). Those studies depict the subjective feeling of intending to act (i.e the *when* and *whether* aspects of intention) as a metacognitive process that integrates both preparatory signals and retrospective elements in a hierarchical organization. Coupled with real-time approach, this theoretical framework allows to inquire "whether and how motor preparation informs the conscious experience of intention" (Pares Pujolras, 2019).

In the present work, we investigate the condition under which people are aware of the content of their decisions. As mentioned in Chapter 1, the decision content (i.e. the alternative chosen by the decision maker) could be encoded at different level of the decision process. First people might be prospectively aware of their intention to choose a given alternative by accessing their action selection process. Yet in the physical world, different decisions always correspond to different motor action. Therefore, people might retrospectively infer the option they unconsciously chosen based on the monitoring of the motor signals executing this choice.

In this chapter, we described how our experimental approach based on a BCI decoding covert, feature based, selective attention can help us to tackle this issue. In a nutshell, our BCI setup allows people to make a decision of preferentially attending one among two items on the computer screen. Those items are presented at the same location to prevent spatial cues and gaze monitoring from participating in the inference processes on decision content. In Section 2.1, we detail the characteristics of covert, feature-based selective visual attention that bring us to specifically target this cognitive process for our BCI. As a direct decoding approach of selective attention might be challeng-

ing, we present in Section 2.2 the steady state visual evoked potential methods that allow to measure consequences of attention allocation on the visual system. Finally in Section 2.3 , we describe the theoretical concept and the mathematical procedure underlying our BCI setup.

2.1 SELECTIVE VISUAL ATTENTION IN THE BRAIN

Although cognitive paradigm emphasized the analogy between brain and computer (Searle, 1990b; Searle, 1990a), brain computation has to minimize the number of spikes needed to transmit a given signal(Barlow, 1961). Indeed, neural activity, and more precisely neural firing rates account for the larger part of brain metabolic cost (Attwell and Laughlin, 2001). As we have seen in Chapter 1, one way to tackle complex environments at reduced cost is to adopt a predictive coding approach, where only the difference between brain prediction and actual sensory inputs travel up the cortical hierarchy. In addition, structural solution such as sparse coding have been described in the brain (Chalk, Marre, and Tkačik, 2018; Perez-Orive et al., 2002; Vinje and Gallant, 2000; Harris and Mrsic-Flogel, 2013; Perez-Orive et al., 2002; Vinje and Gallant, 2000; Harris and Mrsic-Flogel, 2013). Indeed, the sparseness of neural coding ranges on a continuum from dense, highly redundant representation, to local representation encoded in one single neuron (Barlow, 1969; Quiroga et al., 2005). Recent studies reveal that different coding schemes meet different computational requests (Raymond and Medina, 2018) like predicting the future or encoding past events.

Above its account by predictive coding theory (Friston, 2012; Hohwy, 2013; Clark, 2015), attention can be primarily described as a mechanism contributing to efficient coding. Indeed, neuron spikes metabolic cost imposes a drastic limitation to the number of neurons that can be simultaneously activated (Lennie, 2003). Thus attention operates as a mechanism of selective attribution of computational resources. From a bottom up perspective, visual stimuli compete for the access to computational resources such that neurons representing different features of the visual scene engage in competitive interaction, resulting in the suppression of the non selected features (Desimone and Duncan, 1995; Carrasco, 2011). From a top down perspective, attention selectively allocates available computational resources at the attended features to the detriment of the unattended ones (Carrasco, 2011).

Although both perspectives reflect the limitation of the brain computational ability, they also illustrate different aspects of selective pressure exerted on the brain by its environement. In fact, bottom-up perspective corresponds to the need for the brain to stick with environmental dynamics while top-down perspective reflects the necessity for the agents to comply with their current goals. In line with

this disjunction, neuroimaging studies described three networks underlying different aspects of selective attention mechanism (Posner and Petersen, 1990).

2.1.1 *Neuro-anatomy of visual selective attentional process*

According to Posner and Petersen, 1990, the attentional system fulfill three main functions that have been associated to three distinct brain networks (Posner et al., 1988; Posner and Petersen, 1990; Carrasco, 2011; Petersen and Posner, 2012):

- *The orienting system* prioritizes sensory inputs by selecting a modality or location via either "foveating" specific stimulus or covertly attending it. Thereby, computation of the attended stimulus is improved. Orienting function has been associated with posterior brain areas including the superior parietal lobe, the temporal parietal junction and the frontal eye fields .
- *The executive control system* regroups top down executive control networks involved in conflict monitoring and top down control. Anatomically, it has been associated with an anterior attentional system, which involves the anterior cingulate and the lateralprefrontal cortex.
- *The alerting system* maintains a certain degree of sensitivity or arousal to incoming stimuli. This modulation of general excitability is controlled by a network spanning over the frontal and parietal regions of the right hemisphere.

On top of this neuro-anatomical description, psychophysics and cognitive psychology studies classified selective attention along three main axis (Moore and Zirnsak, 2017). Although these forms of attention might be similar at certain mechanistic level (e.g. Hohwy, 2013), it can be relevant to separate them experimentally as can have different effects on the rest of cognition.

- *Spatial versus Feature-based attention* Attention can be directed toward spatial location of the environment (e.g. stimulus on the right visual field) or directed toward a specific feature (e.g. blue stimulus versus red stimulus).
- *Overt versus Covert attention* Attention can selectively process a stimulus in absence of any orienting movement toward this stimulus. It is distinguished from a selective process accompanied by orientation (e.g. of the gaze) toward the selected stimulus location.
- *Top-down versus Bottom-up attention* Attention could be an endogenously generated process, reflecting motivation, preferences or strategy. Alternatively attention could be an exogenously driven process whose selection is solely based on the salience of stimuli.

As we mentioned before, we address in the present work whether and how decision processes inform participants about their choices. Therefore, participants have to make decisions in a situation in which we can control or measure all the cues that could potentially notify them about their choice. To address this issue, participants make a choice which consists in preferentially attending one among two overlapping items presented at the center of the screen. This decision involves top-down, feature based covert attention and avoids cues like eye movement or spatial location to grant participants with information about their choice.

2.1.2 *Feature-based covert attention (FBCA)*

FBCA has been shown to increase visually driven fire rates encoding the attended stimulus along the all visual path, beginning as soon as the thalamic relay of visual stream, namely the dorsal lateral geniculate nucleus (Saenz, Buracas, and Boynton, 2002). FBCA not only increases the strength of spiking activity but also the information those spikes convey at both single neuron and population scale. In addition, attention could be linked with increased synchrony among the neurons and neuron populations coding for the attended stimulus (Moore and Zirnsak, 2017).

From a behavioral perspective, knowing in advance a feature of the target stimulus improves participants' detection performance. Moreover, such pre-cueing can affect low level sensitivity like motion detection or direction discrimination. Indeed when cued with the direction of moving dots, participants are more sensitive in motion detection. Importantly, improvements are still effective in the presence of distractor. This could indicate that FBCA specifically boost the behaviorally relevant psychophysical channel (Carrasco, 2011) and thereby reflect the cue driven decision process.

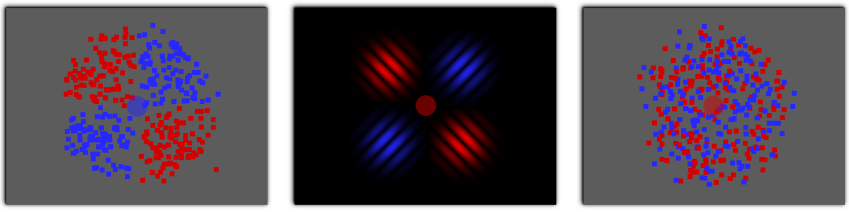

Figure 3: Example of visual stimulation targeting FBCA.
In the three examples shown here, participants were asked to attend either red or blue dots while their gaze remained located on the central cue.
Data presented here are extracted from our pilot study inspired by Müller et al., 2006 work.

2.1.3 *Selective attention and decision*

As described by the predictive coding theory, attention filters behaviorally relevant information by modulating their expected precision (see Section 1.2). Indeed attention favors processing of relevant information and allows the attenuation of distractors (Kinchla, 1992; Carrasco, 2011). Furthermore, motivation and reward alter attention allocation (Bourgeois, Chelazzi, and Vuilleumier, 2016). Drift diffusion models of decision posit that decisions are based on evidence accumulation operated during gaze fixation (Krajbich and Rangel, 2011). Yet, attention should not be reduced to a passive information sampling process, as it rather plays an active role in decision. Attention is not solely guided by top down preferences but also reflects bottom up process and interaction with working memory. Attention is known to impact value integration and confidence during decision (Kunar et al., 2017; Kurtz et al., 2017). Thereby, decision emerges from an interaction between preference, external stimulation, attentional and memory processes (for a detailed review of influence of attention on decision see Orquin and Loose, 2013).

Recent neuroimaging studies began to dissect the complex interconnexion of higher order cognitive function involved in decision processes. Indeed, the relationship between executive control of attention and the decision based processing of information have been associated with inter-connected sub-region of the pre-frontal cortex (PFC) (Hunt et al., 2018). These sub-region are thought to concomitantly process information and guide its sampling by attention shift. In line with those results, sub-region of the PFC reflecting categorical decisions about ambiguous stimuli also drive modulation of visual attention (Roy, Buschman, and Miller, 2014).

Altogether, these studies show the tight relationship between selective attention and decision processes. Attention does not only provide indication about a decision related top-down guidance in information sampling. Instead, it constitutes a core element of the decision process and interact with working memory and valuation process through pre-frontal parallel networks.

2.2 MEASURING THE CONSEQUENCE OF SELECTIVE ATTENTION.

As we have seen, selective attention recruits an ensemble of neural networks spreading through the entire cortex. It is noteworthy that many of the studies having identified those selective attention related networks rely on fMRI measures and non human primate studies. This method achieves a good spatial resolution and provides access to deeper cortical layers at the cost of poor temporal resolution and invasive procedure. Brain computer interfaces on the other hand require fast and reliable decoding of a neural feature. Since frontal areas are highly integrated regions dealing with a wide range of functions,

decoding the attentional process directly might prove unreliable. We thus opted for a method assessing the consequences of selective attention on an evoked visual feature: the steady states visual evoked potential.

2.2.1 *The steady-state visual evoked potential (ssVEP)*

Visual evoked potentials are stereotypical responses to transient visual stimulation, usually assessed through EEG or MEG (magnetoencephalography). If the stimulation is periodically modulated as a function of time, the evoked response obtain by averaging over trials with time locked to the stimulation will be a visual evoked potential of small amplitude termed "steady-state" visual evoked potential (ssVEP) (Adrian and Matthews, 1934; Regan, 1966). Therefore, it is possible to define ssVEP as follows:

ssVEPs are evoked responses induced by flickering visual stimuli. ssVEPs are periodic, with a fundamental frequency corresponding to the visual stimulation frequency. The spectrum of ssVEP is characterized by peaks at stimulation frequencies and related harmonics and is stable over time. (Vialatte et al., 2010).

The brain areas generating ssVEP depends on the frequency of stimulation: high frequencies(> 10-12 Hz) evoked ssVEP sources are localized mostly on V1, lower frequencies evoked ssVEP are likely to originate from pre-cortical structure including retina and lateral geniculate nucleus (Vialatte et al., 2010). If a consensus about the complexity of ssVEP localization has yet to be found, it appears that V1 exhibits the strongest signals across a wide range of stimulation frequencies and visual features (Herrmann, 2001; Vialatte et al., 2010; Norcia et al., 2015). On the mechanistic aspect, ssVEP have been associated with non-linear propagation of visual information through the visual pathway. From non linearity stem the presence of harmonics in the ssVEP spectrum even in absence of those harmonics in the visual stimulation (Labecki et al., 2016). Furthermore, inter-modulation frequencies are found in the ssVEP spectrum when two or more visual stimulations at distinct frequencies are concomitantly presented to participants.

ssVEP have been found to be less sensitive to artifacts like eye movements and blinks, but also to electromyographic noise contamination compared with regular transient visual evoked potential. Since the late 60's, ssVEP have been applied to the study of a wide range of phenomena spanning over both clinical research and cognitive science. Yet in the next section we will focus on the relationship between selective visual attention and ssVEP.

2.2.2 *Visual selective attention modulates ssVEP*

Although ssVEPs has primarily been viewed as a tool to study early cortical stages of sensory processes, they have since been successfully used to explore higher order processes. As the measure they provide is directly attributable to one specific visual stimulus on the screen and presents a high signal to noise ratio (SNR), ssVEPs are particularly adapted to the study of attention (Norcia et al., 2015). In a previous study, participants were presented with two spatially separated letters flickering at distinct frequencies. It results that attending to one letter increases the ssVEP amplitude corresponding to the letter oscillation frequency (Morgan, Hansen, and Hillyard, 1996).

ssVEP can also be used when two stimuli are present at the same spatial location. Feature-based selective attention was first addressed by Pei, Pettet, and Norcia, 2002 in a study where horizontal and vertical bars of 16 cross were oscillating at different frequencies. The study showed that attention modulated neural processing in a spatial non specific manner. Indeed, the ssVEP amplitude was larger at the frequency corresponding to the attended orientation. Dissociating further the contribution of spatial and feature-based selective attention, Müller et al., 2006 used a stimulus consisting in two superimposed random dots kinematograms of distinct colors. The study confirms that covert feature-based selective attention selectively increases ssVEP amplitude of the attended item. Notably, this increase was found to be maximal over occipital area of V1-V3 . Follow up studies dissected the temporal dynamic of the ssVEP modulation by selective attention. Cues were displayed on the screen asking participants to preferentially attend one colored group of dots. The enhancement of ssVEPs amplitude occurs around 220 ms after the cue appearance and is accompanied about 140 ms later by a decrease of the ssVEP amplitude related to unattended dots (Andersen and Müller, 2010).

If ssVEP can reliably track the identity of the currently attended features, whether they can provide information on the quantity of allocated attentional resources is not clear. Clinical studies established that schizophrenic patients have reliably smaller ssVEP during passive exposition but show an hyper-activation compared to control subjects when actively attending a stimulation in the 1-30 Hz range. Schizophrenic patients are known to endure attentional deficits since the early stage of the disease (McGhie and Chapman, 1961). Such deficits include global enhancement of background noise leading to difficulty to selectively allocate attention to relevant part of the environment. Thus, compensatory over-allocation of attention are often observed. This compensation effect echo Bayesian account of schizophrenia according to which schizophrenic patient might fail to attenuate sensory precision of incoming perceptual inputs, leading to compensatory increase in the precision of high-level prior beliefs (Fletcher and Frith, 2009). Similarly, children having lived febrile seizure episode

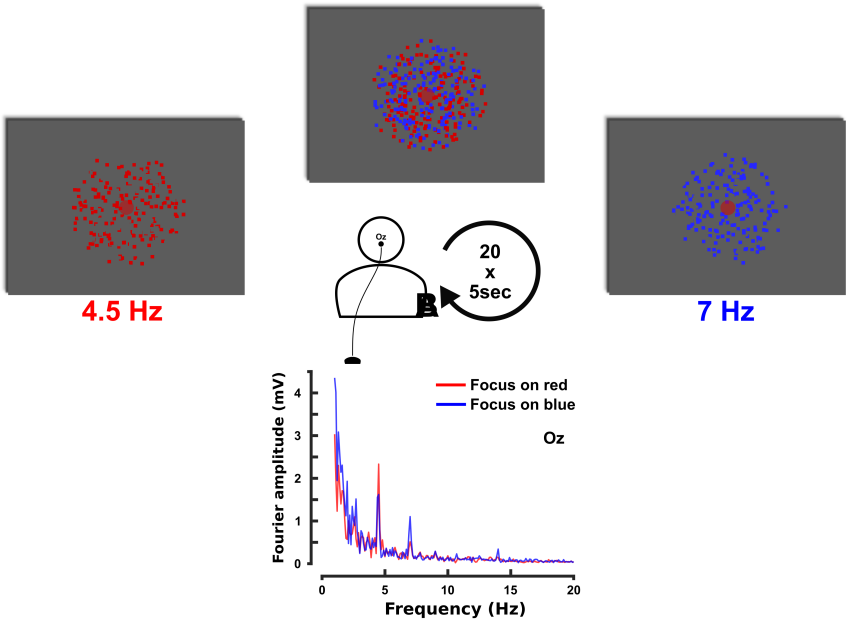

Figure 4: Attention allocation modulates ssVEP amplitude.
A) We presented two superimposed kinematograms of red and blue dot flickering at respectively 4.5 and 7 Hz. Participants were asked to preferentially attend one of the kinematogram for 5 seconds. The color of the target kinematogram was specified by the central dot on which participants were asked to maintain their gaze. B) Allocating attention to one color enhances the amplitude of the Fourier component associated with the corresponding frequency. Fourier transform is presented for one participant average over 20 trials calculated at the electrode Oz. Peak at 4.5Hz and 7Hz are enhanced when participants focus on red or blue kinematogram respectively.
Data presented here are extracted from our pilot study inspired by Müller et al., 2006 work.

show a superior control over working memory, distraction avoidance and attention (Chang et al., 2000; Ku et al., 2014). In parallel, their ssVEP amplitude is larger than age-matched control across the all frequency range (Vialatte et al., 2010). Altogether, those results link the quantitative variation of attention allocation with variation in ssVEP amplitude. Therefore, a ssVEP based decoding approach would be able not only to infer the attended object on the screen but also to track the strength of allocated attention.
Although its use has mostly been circumscribed to the study of visual information processing and attention, ssVEP have recently been employed in paradigm addressing other integrated cognitive mechanisms.

2.2.3 *ssVEP to investigate cognition.*

The effect of attention on ssVEP amplitude predicts behavioral output. As an example in Andersen and Müller, 2010, the lag between

the cue indicating the color of the target dots and the ssVEP amplitude enhancement was strongly correlated with the response time to detect a transient coherent movement in the target dots. Similarly, the performance in an orientation discrimination task was predicted by the orientation-selective ssVEP response (Garcia, Srinivasan, and Serences, 2013).

Recent studies have used ssVEP to differentially tag different levels of visual hierarchical perception (Gordon et al., 2017; Gordon et al., 2019). They combined semantic wavelet induced frequency tagging (SWIFT) (Koenig-Robert and VanRullen, 2013) with classical ssVEP (Norcia et al., 2015) to disentangle bottom up from top down processing. Indeed, ssVEP predominantly tag low level sensory inputs in the visual hierarchy while SWIFT method has been shown to reflect high level top-down processing. They track the degree of integration between top down predictions and bottom-up sensory signals through inter-modulation components of the EEG spectrum (frequencies corresponding to linear combination of SWIFT and ssVEP frequencies). They showed that the participation of top-down predictions[1] decreased as reliability of sensory inputs increase, lending support to predictive coding theory.

In summary, ssVEP analysis provides an indirect measure of several aspects of visual selective attention. Indeed ssVEP grants access to qualitative aspects of selective attention such as the location of the attentional spot or the identity of the preferentially attended object. Furthermore, ssVEP also gives insight about quantitative aspects of the attentional process like the strength of allocated attention or its timing. Recent studies have demonstrated that appropriate stimulation allows us to use of ssVEP to investigate a wide range of cognitive mechanisms. In the next section, we will describe how we implement a BCI exploiting ssVEP to continuously track selective attention and thereby the decision of our participants.

2.3 BRAIN COMPUTER INTERFACE (BCI): TOWARD A REAL-TIME MEASURE OF SELECTIVE ATTENTION

To continuously track the content of the decision of our participants, we used a BCI combining stimulus reconstruction approach with sweep-ssVEP (Regan, 1973; Ales et al., 2012). Stimulus reconstruction has been successfully used to decode the attended stimulation in auditory (O'Sullivan et al., 2015) and visual paradigm (Sprague, Saproo, and Serences, 2015). Although this method achieves a good accuracy, decoding attention toward natural stimulus requires a fair amount of data. To continuously track attention, we follow the most recent development in the field of BCI (Rezeika et al., 2018; Chen et al., 2015) that apply stimulus reconstruction to a simple pattern of oscillation. We describe our stimulation in detail in the next chapter. In the following

1 indexed by SWIFT signal to noise ratio

section, we expose BCI's general concept for decoding participants' decision content.

2.3.1 *Linear stimulus reconstruction and individual model creation*

As we have previously seen, selective attention specifically alters the attended stimulus processing. One way to infer which stimulus among many is attended is to assess those processing modulations via a stimulus-reconstruction approach. This method attempts to reconstruct an estimate of the input by combining the neural signals recorded during stimulus presentation according to a participant-specific model. After the reconstruction steps, the reconstructed signal is correlated with each stimulus oscillations pattern and the correlation factors can be compared to infer the attended stimulus (O'Sullivan et al., 2015).

The reconstructed stimulus is a linear combination of the neural signals received from the electrodes (in the case of an EEG based approach) according to a participant's specific model. To build the individual model, we rely on labeled trials (e.g. trials where the target stimulus is known). We detailed the method in the methodological section of Chapter 3. In a nutshell, the model minimizes the difference between on one hand the combination between the EEG electrodes and on the other hand the target stimulus oscillations pattern.

Applying this method to detect the attended audio stream in a cocktail party results in robust accuracy on a single trial level (O'Sullivan et al., 2015). It includes all the scalp information and thus encompasses top-down and bottom-up contribution to attended stimulus processing. Furthermore, as the model tends to attribute a very low weight to irrelevant information, it does not necessitate advanced methods of filtering and can be conducted without removing blink, muscle or electrical artifact, or even without average-referencing data.

However, studies achieving a good (>80%) classification accuracy rely on a large amount of data. For example, O'Sullivan et al., 2015 classifies 1 minute of auditory data per trial. To gain a better insight in the temporal dynamic of participant decisions, one needs a finer temporal grain in the decoding approach. In line with innovative work in the domain of BCI based speller (Chen et al., 2015), we drastically reduced the required amount of data to 3 seconds by applying the stimulus reconstruction approach to a simple modulation pattern. This procedure allows us to modulate the feedback according to participants' decision to attend one stimulus among two in 5 seconds long trial.

2.3.2 *The BCI loop*

Brain Computer Interface is a recent neuro-technology which development is giving rise to a flourishing field of research, both in the

engineering and fundamental aspects. Although its first description can be traced back in the 70's (Vidal, 1973; Vidal, 1977), BCI setup long remained mere proof of concept and concrete application for general public use began only recently to be extensively developed (McFarland and Wolpaw, 2017). BCI setups have been developed with just about every neuro-imaging method (invasive or not) existing, connecting vision but also sensory-motor system (Wolpaw et al., 1991), audition (Klobassa et al., 2009) and even taste (Canna et al., 2019). In the present work, we focus on EEG-based visual BCI.
Among its several applications, the main effort has been put in restoring communication and control in paralyzed patients (McFarland et al., 2017). Derived from this line of research, BCI based spellers allow people to write by only attending a virtual keyboard (Rezeika et al., 2018). Other domains of application like video game control are progressively emerging, building on methodological breakthroughs of their predecessors (Kerous, Skola, and Liarokapis, 2018).

Pro-active on this domain, the engineer literature usually define 4 modular subsystems in a BCI (Wolpaw and Wolpaw, 2012; McFarland and Wolpaw, 2017):

- *Stimulation:* Here visual, the stimulation elicits an appropriate neural pattern to be decoded. In Appendix A, we detail our method to present participants with two equivalent choice options distinguishable patterns.

- *Signal recording:* The neuro-imaging method used to continuously extract brain signals. In the present work, we used a 64 electrodes electroencephalography.

- *Signal processing:* This subsystem regroups the processing of the raw brain signals along with the extraction of features relevant for BCI mediated interaction. We described the detailed procedure in the Methodology section of Chapter 3. Our goal is to maintain a robust and accurate classification while capturing the dynamic of selective attention allocation. We thus applied a stimulus reconstruction approach on a sliding window of 3 seconds. Each 250 milliseconds, we computed the reconstructed stimulus and correlated it with the pattern of oscillation of our two potential targets.

- *Interaction:* Finally, the fourth subsystem specifies system operation. It notably entails how, from the relevant features extracted by the *signal processing* subsystem (in our case the results of correlation between reconstructed signal and display stimuli pattern), the BCI decides which item was attended. Thereby,BCI adapts the feedback to the participant's neural signals.

Below is illustrated our implementation of BCI subsystems allowing us to investigate the simple decision of preferentially attending one over two items overlapping in the screen center.

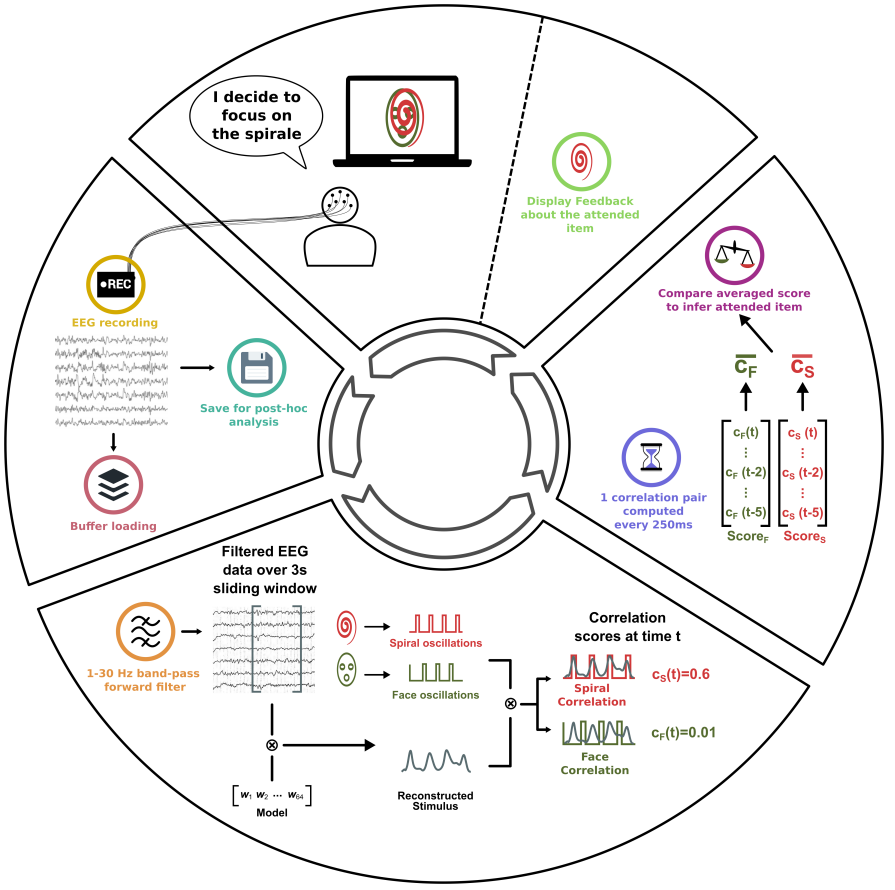

Figure 5: Brain Computer Interface principle.
The four subsystems of a BCI are detailled in Section 2.3.2. Top panel) Stimulation. Left panel)Signal recording. Bottom panel) Signal processing. Right panel) Interaction implementation.

2.3.3 *Summary*

To conclude, attention is an omnipresent cognitive mechanism intermingled with virtually all other brain mechanisms. Especially relevant for the present work, participants' decisions influence covert feature based attention allocation. Therefore, to track decisions in time, we choose to rely on a BCI approach to decode selective attention allocated to one among two overlapping items. We present in Appendix A the visual stimulation we developed for our BCI allowing participants to perform a voluntary covert decision. Although such a picking choices lack the motivational aspect that characterize real-life decisions, they are thought to rely on similar mechanism. Furthermore, a non-ecological approach allows us to specifically target proximal aspect of decision while neither long term planning nor motor preparation can be informative for participant awareness. Thus, we can address whether and how the representation available at this specific level are susceptible to enter consciousness.

2.3.4 *Empirical contribution overview*

We then present 2 experimental contributions on decision content awareness, using covert attention-based Brain Computer Interface approach.

In Chapter 3, we directly address the details of endogenous and exogenous contribution to people's introspection about their recent decision. We used BCI to continuously measure endogenous variables and control external feedback during participants' decisions. We then probe participant's recent decision awareness and compare their answer to the decision inferred by our stimulus reconstruction approach. This first study provides evidence that introspection results from a Bayesian integration process involving both external feedback and internal decision variables.

Then in Chapter 4, we go further by exploring whether and how people could be prospectively conscious and in control of the content of their impending decision. We develop a new approach based on the combination of stimulus reconstruction to track ongoing deliberation with a process dissociation procedure (PDP) (Jacoby, 1991). In PDP, participants try to avoid using their early deliberation (exclusion task) or make sure they do use it (inclusion task) when asked to make a decision. This second study demonstrates that before decision execution, participants can be aware of an ongoing deliberation process but can not access the option selected during the deliberation.

Part II

EXPERIMENTAL CONTRIBUTIONS

3

IMBALANCED EXOGENOUS AND ENDOGENOUS CONTRIBUTION TO INTROSPECTION LEAD TO ILLUSION AND ASSOCIATED METACOGNITIVE FAILURES.

3.1 SYNOPSIS

People can introspect on their internal state and report the reasons driving their decisions but Choice Blindness (CB) experiments suggest that this ability can sometimes be a retrospective illusion. Indeed, when presented with deceptive cues, people justify choices they did not make in the first place, suggesting that external cues largely contribute to introspective processes. Yet, it remains unclear what are the respective contributions of external cues and internal decision variables in forming introspective report. Here, using a brain-computer interface, we show that internal variables continue to be monitored but are integrated and dominated by deceptive cues during CB episodes. Moreover, we show that deceptive cues overturn the classical relationship between confidence and accuracy: introspective failures are associated with higher confidence than accurate introspective reports. We tracked back the origin of these overconfident confabulations by revealing their prominence when internal decision evidence is weak and variable. Thus, introspection is neither a direct reading of internal variables nor a mere retrospective illusion, but rather reflects the integration of internal decision evidence and external cues, with CB being a special instance where internal evidence is inconsistent.

3.2 INTRODUCTION

Humans constantly monitor their choices and actions to adapt their behavior (Ericson and Simon, 1980; Ridderinkhof et al., 2004; Ullsperger and Von Cramon, 2004). This ability typically involves introspective mechanisms that are used to evaluate and justify decisions (Moshman, 2014). Yet, introspection turns out to be unreliable on many occasions (Nisbett and Wilson, 1977). For instance, participants can believe they have intentionally performed an action that was actually initiated by another agent (Wegner and Wheatley, 1999; Wegner, 2002). Similarly, participants can confabulate about why they choose an option while they actually made the opposite choice in the first place (Johansson et al., 2005; Johansson et al., 2006; Hall et al., 2010). A striking example of introspective illusion is given by choice blindness (CB) experiments. In this paradigm, participants select which one of two faces is more attractive, and are then presented with the option they selected and asked to justify their decision. On some trials, they

are lured to have chosen the non-preferred face. Yet, they provide confabulated justifications about why this face is more attractive than the other. This phenomenon, which has been extended to economic decisions, political preferences and moral judgments, reveals that introspection can be, under certain circumstances, a retrospective illusion (Hall, Johansson, and Strandberg, 2012; Hall et al., 2013; Hall et al., 2013; McLaughlin and Somerville, 2013; Strandberg et al., 2018).

Yet, participants have also been shown to have reasonable introspective access to the elements driving their decision (Ericson and Simon, 1980; Grover, 1982; Schultze-Kraft et al., 2016; Reyes et al., 2018; Parés-Pujolràs et al., 2019). These apparently contradicting results could be reconciliated under an integrative account of introspection where both internal decision variables and external, contextual cues contribute to participants' introspective reports. Yet, although such view began to receive empirical support (Schultze-Kraft et al., 2020), the modalities under which these two components could be integrated during introspective processes remained unsettled.

One way to investigate the formation of introspection about decisions consists in studying how internal decision variables impact CB episodes. In line with previous Bayesian integrative accounts of introspective processes (Moore and Fletcher, 2012; Legaspi and Toyoizumi, 2019), we predicted that impact of internal decision evidence on introspection would be mediated by their availability and reliability. Furthermore, the integrative process would modulate not only the quantity of introspective failures (i.e., the amount of CB episodes), but also their quality (i.e., how much participants are convinced in their confabulation). That is, when a reliable source of external cue sometimes provides a deceptive information, participants would confabulate with high confidence (dominated by external cues) when their internal decision evidence is weaker.

Here, to address this issue, we relied on a brain-computer interface (BCI) setup to track participants' internal decision variables during the original choice (i.e., prior to the external cue and report). Participants had to freely choose to preferentially attend for 5s to one out of two overlapping stimuli while their EEG was recorded and a marker of selective attention was measured in real-time (decision phase, see Figure 1A). Recent neuroimaging studies revealed that top-down attentional mechanism reflect decision processes (Gottlieb and Balan, 2010; Roy, Buschman, and Miller, 2014; Hunt et al., 2018). Although we did not measure decision mechanisms per se, our neural index allowed us to track their consequence in the form of selective attention allocation to one category over the other. This BCI setup allowed not only to measure a proxy of internal decision evidence independently of introspective reports but also to control de reliability of external cues. Following the decision phase, participants were presented with a feedback cue that matched their original choice in 75% of the trials (informative feedback). Importantly, they were presented with the alternative, non-preferred choice as the outcome of their decision in 25% of the trials (deceptive feedback). Moreover, in order to assess the impact of internal decision variables not only on the quantity of in-

trospective failures but also on their quality, participants were asked to rate the confidence they experienced in their decision.

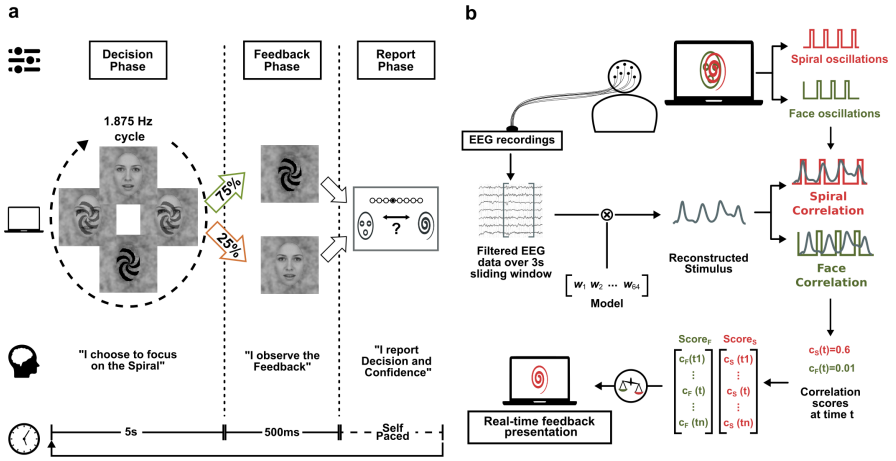

Figure 6: A) Experimental paradigm. Each trial comprised three phases. 1) Decision phase: participants were presented with overlapping face and spiral oscillating at 1.875Hz in temporal phase opposition and asked to choose one or the other category and focus on it until the end of the 5s period . 2) Feedback phase: participants were then presented with a feedback cue for 500 ms, reflecting their recent decision on 75% of the trials (green, informative trial) or the opposite choice in the 25% remaining trials (orange, deceptive trial). 3) Report phase: participants were then requested to report the object they preferentially attended before the feedback cue along with their confidence in this report on a 4 steps scale.
B) Real-time decoding procedure. Each 250 ms, a reconstructed stimulus was computed by linearly combining the 64 EEG electrodes signal over a 3s window according to the model weights computed beforehand. We then obtain correlation scores for both face (green) and spiral (red) stimuli by computing the correlation between the reconstructed stimulus and the expected face and spiral oscillations respectively. On the end of each trial, correlation scores computed during the last 1.5 seconds were averaged separately for each category, and the highest average was considered the attended category for the presentation of the feedback cue.

3.3 RESULTS

3.3.1 *Impact of internal decision evidence and external cues on introspection.*

Does introspection integrate internal evidence supporting just-made decisions or is it a pure reconstructive process shaped by external cueing? We here operationalize introspective accuracy as being correct when the reported object matches the object decoded by the BCI (e.g., for instance face reported, face decoded), and incorrect otherwise (e.g., face reported, spirale decoded). Together with this measure of introspective accuracy, we computed a proxy for internal evidence

accumulation during the decision phase called internal evidence (IE). During the decision phase, our BCI continuously outputs correlation scores associated with each object (Figure 6B). These correlation scores reflect how close are the brain signals from being generated by the observation of the face or the spiral respectively. Then, IE consists in the accumulated difference between the correlation scores associated with each object over the last 3 seconds of the decision phase, and provides an objective measure of how strongly participants preferentially attended one object over the other one (See Methods). Finally, the type of feedback cue displayed on each trial was encoded as either deceptive (opposite to IE in 25% of the trials) or informative (corroborating the IE in 75% of the trials). To determine the respective influences of internal decision evidence and external information on introspective processes, we modelled accuracy using IE and the type of feedback cue as fixed effect and participants as random effect.

We first thought to determine the relationship between the internal evidence that was available during the decision phase and the accuracy of introspective reports. As shown in Figure 7A, introspective accuracy significantly increases with the amount of internal evidence (generalized linear mixed effect model (GLME): Odds Ratio (OR)=1.33, confidence interval (CI)=[1.26 1.41], χ^2 = 211.5, p<0.001). Moreover, the type of feedback had a significant influence on introspective reports as we observed a higher accuracy following an informative feedback (M=0.71, SD=0.11), compared with a deceptive feedback (M=0.57,SD=0.16). (Figure 7B, Section B.3 Table 1) (OR=1.72, CI=[1.45 2.04], χ^2 =151.2, p<0.0001), revealing that the contextual cue influences introspective reports. We found no significant interaction between IE and the type of feedback (OR=1,04, CI=[0,98-1,17], χ^2 = 2.2, p>0.1), revealing that feedback cues modulate the accuracy of introspective reports regardless of the internal information available during the decision phase (Figure 7C).

Is introspection impaired by deceptive feedback or improved by informative external cues? To address this question, we ran an additional control experiment for 16 out of the 30 participants. They performed an additional block with the exact same structure except that no feedback cue was now presented between the decision and report phase. The results of this control session confirm that informative feedback presentation increases the accuracy of participants' reports compared with reports without feedback (t(15) = 2.60, corrected p-value < 0.05). Conversely, we observed that accuracy decreases following a deceptive feedback compared to a condition without feedback (t(29) = 1.96, corrected p-value = 0.069; see Figure 7D and Section B.1). Together, these results reveal that both internal decision evidence and external cues influence participants' introspective reports.

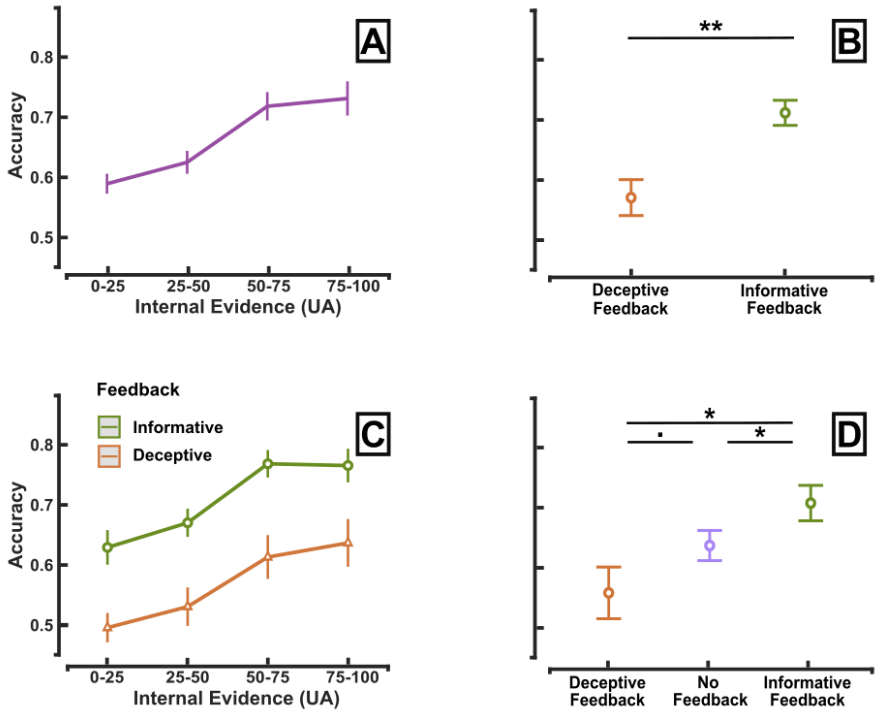

Figure 7: (A-C) Impact of Internal Evidence (IE) on the accuracy of introspective reports. For each participant, we computed the distribution of internal evidence across trials in terms of percentile. Vertical bars represent bootstrapped confidence intervals across participants (1000 iterations). (B-D) Impact of the feedback cue on the accuracy of introspective reports.

3.3.2 *Metacognitive failures:*

We then studied whether external cueing impacts not only reports, but also decision confidence. Confidence is known to track performance for decisions conducted under perceptual uncertainty (Yeung and Summerfield, 2012; Pouget, Drugowitsch, and Kepecs, 2016). Therefore, we aimed at investigating the relationship between introspection and confidence, and in particular whether the misleading influence of the deceptive feedback would also impact the confidence associated with introspective reports. We thus modeled introsepction accuracy using confidence and the type of the feedback as fixed effect and participants as random effect.

We found that when participants are presented with a deceptive feedback cue, the classical relationship between confidence and accuracy is overturned. We observed a significant interaction between the nature of the feedback and the confidence attributed to decision reports (GLME: OR= 3.46, CI= [2.29-4.01], χ^2 =274.4, p<0.001, See Section B.3, Table 1) (Figure 8A). Indeed, when participants received an informative feedback, we found a positive correlation between confidence and introspective accuracy (high confidence: M = 0.80, SEM = 0.02; low confidence: M = 0.58, SEM = 0.03, z(29)=5.2,

d=1.53, p<0.0001 signed-rank test). However, strikingly, when participants received a deceptive feedback, this correlation was inverted, with confidence rising up as accuracy decreased (high confidence: M=0.46, SEM=0.04; low confidence: M = 0.66, SEM = 0.03, z=-4,19, d=-1.07, p<0.0001 signed-rank test). Moreover, we found that participants exhibit a lower confidence after detecting a deceptive feedback (M=1.39 SEM=0.04) than when the deceptive feedback went undetected (M=1.58, SEM=0.04) (signed rank test, z(29)=4.6, d=0.93, p<0.0001, see Figure 21A). Therefore, participants show higher confidence for asserted confabulation than for genuine introspection. Together, these results reveal that a deceptive feedback can not only delude participants about the choices they made (i.e., choice blindness) but also falsify the feeling of confidence they associate with their introspection (i.e., aberrant metacognitive failures).

3.3.3 *Reliability of internal decision evidence:*

How external cues can impact qualitative aspect of introspection such as the confidence? To better understand the underlying mechanisms of this overconfident confabulations, we investigated the conditions that permit exeternal cues to dominate introspective reports. The human brain constantly integrates information coming from multiple noisy sources. Several models propose that in multi-sensory perception, the respective participation of each source of evidence to the final percept is regulated by their own strength and reliability (Ernst and Banks, 2002; Knill and Pouget, 2004; Moore and Fletcher, 2012). Here, we propose that a similar mechanism operates between internal decision evidence and external cues in the production of introspective reports. Under this hypothesis, introspective processes should be dominated by external cues if the internal decision evidence are inconsistent (i.e., weak and noisy).

We built an index of internal consistency acounting for the strength and reliability of internal decision evidence during each trial. To compute internal consistency, we took for each trial the ratio of the internal decision evidence strength divided by its variance (see Methods). Then, to understand how internal decision consistency evolve during confident confabulations, we modeled internal consistency using accuracy and confidence as fixed effect and participants as random effect. To distinguish confabulations from correct introspections, this analysis was restricted to deceptive trials (see Section B.1). We predicted that confabulations associated with high confidence correspond to decisions supported by low internal consistency. In such cases, external cues should prevail in the formation of introspective reports, both in the reported decision and in the associated confidence.

Consistency of internal evidence was higher during correct internal monitoring than for confabulations regardless of confidence (differ-

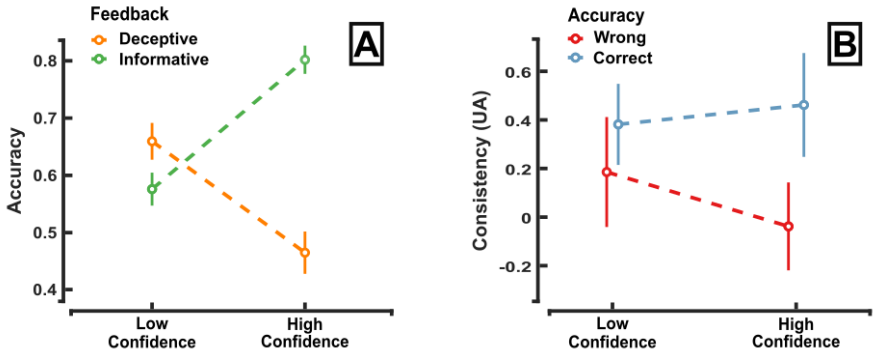

Figure 8: (A) Effect of Feedback on the Accuracy-Confidence relationship. Accuracy in y-axis is the percentage of correct trials and was computer for trial associated with respectively a low and a high confidence separately for deceptive (orange) and informative (green) freedback. (B) Effect of internal decision evidence consistency on confidence in confabulation and accurate introspective reports. Internal Evidence consistency in y-axis was averaged across trials grouped by confidence and accuracy. Only trials followed by deceptive feedback are represented as this condition allows to disentangle correct introspection from confabulation.

ence between correct and incorrect trial, signed rank test: Low confidence: z=2.4, d=0.6 p<0.05, High confidence: z=5.2, d=1.4, corrected p-value<0.001). Furhtermore, we observed a different relationship between confidence and internal consistency for confabulation and accurate introspection (interaction between confidence and accuracy, LME: Estimate=0,18, CI=[0,09 0,3], χ^2 = 8.2, p<0,01), see Figure 8B and Section B.3, table 5. Indeed, consistency was inversely correlated with confidence for confabulated reports (Low confidence: M = 0.15, SEM= 0.07; High confidence M=-0.08, SEM= 0.05), signed rank test z(29)=2.4, d=0.51, p<0,05. Conversely, consistency tended to increase with confidence for accurate introspection, although this trend did not reach significance (low confidence: M=0.35 SEM=0.05; high confidence: M=0.43, SEM= 0.06; paired t-test t(29)=-1.32, d=-0.19, p=0,19). These results reveal that confabulation occurs when internal decision evidence are weak and noisy. Furthermore in such condition, external cues influence not only the content but also the metacognitive aspects of introspective reports on decision.

3.4 DISCUSSION

In the present study, we used a Brain-Computer Interface to covertly track a correlate of internal evidence supporting decisions and study how it affects introspective illusions. We found that participants' introspective reports combine both internal decision evidence and external cues about their decisions. When presented with feedback cues that opposed their original internal decision evidence, participants tended to report it as reflecting their own decision. Furthermore, we found that the noisier and weaker was the original internal decision

evidence, the more they were confident about having made the choice corresponding to the external cue. In other words, external cues dominate introspective reports and decision confidence when internal evidence supporting the original decision was inconsistent.

Combining multiple sources of noisy information has been proposed to account for multisensory perception (Ernst and Banks, 2002; Knill and Pouget, 2004) and for the sense of agency (Synofzik, Vosgerau, and Newen, 2008; Moore and Fletcher, 2012; Legaspi and Toyoizumi, 2019). The sense of agency appears to result from an integration of both internal motor signals (Blakemore, Wolpert, and Frith, 2002; Haggard, Clark, and Kalogeras, 2002; Haggard, 2017) and external information (e.g., action outcome) (Moore and Haggard, 2008; Moore, Wegner, and Haggard, 2009). Moreover, this integration appears to follow Bayesian principles, whereby sources of information are weighted by their respective reliability (Moore and Fletcher, 2012; Legaspi and Toyoizumi, 2019). Here, we propose to extend this framework to account for introspective illusions such as CB. In the context of our study, both external cues and internal variables (here indexed by a correlate of internal evidence supporting the decision) can be considered as noisy sources of information with their own relative contributions to introspection. Therefore, we propose that external cues are combined with internal decision evidence in inverse proportion to internal evidence availability and reliability when forming introspective reports (for similar accounts in the perceptual domain, see (Ernst and Banks, 2002; Clark and Yuille, 2013; Farrell and Lewandowsky, 2018). That is, when a choice was made on the basis of weak or unreliable internal evidence, introspective reports are more likely to be dominated by exogenous elements such as the decision's outcome feedback, thereby resulting in a CB episode.

To unravel which factors influence CB, previous studies relied on behavioral measures and compared the detection rate of deceptive trials across various types of stimuli. For instance, decisions involving familiar choices (e.g., known brands, political preferences, etc) are rarely followed by CB episodes (Hall et al., 2010; Hall, Johansson, and Strandberg, 2012; Sauerland et al., 2014; Somerville and McGowan, 2016; Rieznik et al., 2017; Strandberg et al., 2018). Our work offers an interpretation of those findings by suggesting that familiarity with the choices might increase the weight of internal decision evidence during introspection. Consequently, the influence of internal decision evidence on introspection will prevail over the influence of deceptive cues, thus improving introspective accuracy.

Conversely, one might expect a similar effect if, instead of increasing the consistency of internal decision evidence, it was the external outcome consistency that was decreased. For instance, a recent study (Reyes et al., 2018) manipulated the confidence that participants have on the experimenter and showed that they undergo stronger CB effects when the experimenter appears in control of the experimental

setup or if they have been primed about her professionalism. On the other hand, when primed with an apparent lack of competence of the experimenter or if the experimenter looks overwhelmed by a fake bug on the experimental setup, the detection of deceptive trials largely increases.

Other attempts to address the underlying mechanisms of CB phenomena relied on linguistic analysis but failed to differentiate between reports following deceptive versus non-deceptive trials (Johansson et al., 2005; Johansson et al., 2006; Johansson, Hall, and Sikström, 2008). Together with previous studies (Nisbett and Wilson, 1977), these results argue that introspective reports are based on participants' belief about their decision rather than the mental states supporting those decisions. Participants remain ignorant of those underlying mental states even in the absence of deceptive feedback (Petitmengin et al., 2013). Our results corroborate those conclusions by offering a mechanistic account for why no linguistic difference should be observed between reports following deceptive and non-deceptive feedback. Indeed in both types of trial, internal evidence supporting the decision can be weak and noisy, leading subsequent justifications to mostly reflect the feedback presentation rather than internal variables. Therefore, no difference should be expected in the justification of informative and deceptive trials.

While the use of confidence judgment is widespread in psychological studies, its relationship with confabulation is still unclear. While some argue that introspective illusions are subjectively indistinguishable from alleged introspection (Carruthers 2009, 2010), other studies show that illusions often come with a reduced confidence (Nisbett and Wilson, 1977; Wheatley and Haidt, 2005; Hall, Johansson, and Strandberg, 2012; Rieznik et al., 2017; Strandberg et al., 2018). Altogether, our results nuance this debate by showing that the subjective distinction between illusory and alleged introspections depends on the availability and reliability of internal variables. When internal decision variables are weak and noisy, confabulation can't be distinguished from accurate introspection, and both will be reported with high confidence (Carruthers, 2009; Carruthers, 2010). If the consistency of the internal decision evidence is high, participants directly access their recent decision variables and easily detect external manipulations. Finally, if the internal evidence supporting decisions shows intermediate consistency, participants will eventually fail to notice external manipulation but their subjective experience will be affected as they report a lower confidence compared to correct introspections (Fiala, Nichols, and Carruthers, 2009) (see Figure 8B).

Although our task presents many similarities with the original Choice Blindness paradigm, we must note that it also differs on several aspects. We manipulate the feedback more often (random ordering assignment in 25% of the trials) and in a more explicit manner (participants were informed of the potential deceptive nature of the

outcome) compared with the original paradigm. Importantly, however, we still observed a large portion of CB episodes though reduced compared to the original study (40% here versus 60% to 80% in original study (Johansson et al., 2005) (see Figure 21B).

In the present study, we propose participants could introspect some of their internal information but are subject to introspective illusions when those information are weak and noisy. Yet this interpretation rely on the assumption that our BCI decode accurately the object participants choose to focus on. Nonetheless, our BCI could sometimes misclassified participants decision, leading to the presentation of the opposite feedback. The reliability of our decoder approximate 80% during externally driven choice (i.e. the model building). In line with recent findings (Schurger, Sitt, and Dehaene, 2012; Murakami et al., 2014; Wisniewski, Goschke, and Haynes, 2016; Brass, Furstenberg, and Mele, 2019), we argue that performance of the decoder should be similar for voluntary decision. Yet, our decoding method could be sensitive to decision change in the last second of the decision phase. We improve further the robustness of our decoding methods by averaging several BCI outputs together to perform the decision classification. Moreover to account for the potential late change of decision, we first identified trials where those changes potentially occurred and confirmed our results after having removed them (see Appendix B, Materials).

3.4.1 *Conclusion*

In conclusion, we combined a choice blindness paradigm with a brain-computer interface to demonstrate that introspective reports about recent, private decisions result from the integration of internal evidence and external cues. When internal variables supporting the original choice are weak and noisy, participants accept external outcomes as their original intention even when the two are in contradiction. Moreover, our study reveals that not only the object of a decision but also the metacognitive aspects of this decision are subject to reconstruction. When internal decision evidence is weak or unreliable, participants show high confidence for their confabulations. Our study shed new lights on the mechanisms underlying introspective illusions, unraveling a continuum in the awareness people have about their decisions. Indeed, their introspective experience ranges from relying on internal information to being purely driven by external factors as a function of the availability and reliability of the evidence supporting their original decision.

3.5 METHODS

3.5.1 *Participants*

Thirty healthy participants with normal or corrected-to-normal vision took part in the experiment (14 males; all right-handed; mean age: 25.1 years, SEM = 3.4). Two additional participants were tested but were not included in the analysis because of EEG artifacts (N=1) or technical failures (N=1) altering the online experiment. All participants signed a written consent and received financial compensation in exchange for their participation. This experimental protocol was approved by the local ethical committee (Conseil d'évaluation éthique pour les recherches en santé, Paris, France).

Participants performed 480 trials in the main experiment. All trials contained a decision and feedback phase. The feedback was informative in 75% of the trials, but deceptive in the remaining 25%. Introspective reports were required on all deceptive trials, but only on a third of the informative trials, in order to balance them across cue validity. This led to the analysis of 120 deceptive trials and 120 informative trials per participant. In addition, 16 out of the 30 participants underwent a session with a control condition consisting of an extra 160 control trials where no feedback was presented, here again with half the trials including introspective reports. To account for potential order effect, the control session was presented after 1 third, 2 third or at the end of the main experiment.

3.5.2 *Visual stimulation.*

Visual stimuli consisted of the superposition of two half transparent animated images, a face and a spiral, at the center of the screen (Iiyama ProLite E2483HS-B3). The spiral rotated around its center while the face alternatively opened and closed its mouth. Such superposition of half transparent animated streams has been shown to reduce the stability of the percept containing the two streams and thus facilitates the voluntary switch from one item to the other (Neisser and Becklen, 1975; Clark, 2017; Ransom, Fazelpour, and Mole, 2017). In addition, the two animated streams had to evoke distinguishable brain responses in order for our BCI to decode the attentional focus of the participant. We therefore continuously modulated the spatial phase scrambling of each item, eliciting "sweep" steady state visually evoked potential (ssVEP) responses (Ales et al., 2012; Norcia et al., 2015) at the frequency of 1.875 Hz for both streams but in temporal phase opposition.

To build our animated stimuli, we used 12 images of a face regularly spanning an animation of mouth opening and 8 images of a homemade spiral at different steps of a rotation animation. We cropped each image with a gaussian filter to obtain smooth edges.

We then inserted each image in a noisy background with the same Fourier amplitude. As a result, the items of both categories appear to emerge from the noise as phase scrambling decreases (Ales et al., 2012). Then, for each step of animation and each item we selected the correct amount of phase scrambling to produce the desired sweep ssVEP. Spatial phase scrambling was computed by a phase interpolation method (Ales et al., 2012). Finally, we superimposed pairs of images to create a complete dynamic stimulus of two superimposed animated images, producing oscillatory signals at the same frequency but in phase opposition.

In half of the trials, participants were asked to report the object of their decision (i.e., object they had decided to preferentially attend) and how confident they were about this decision. Both the object of decision and their confidence in that decision were reported at the same time on a 4 levels scale. Eight circles were thus displayed on a horizontal line (i.e., 4 circles for each object). A reference central dot was displayed between the two most central circles to ensure forced-choice decisions. Participants reported their decision by choosing to move the dot either to the left or to the right (counterbalanced across trials), and their confidence was rated by choosing a circle that was close (very unconfident) or far (very confident) from the reference central dot (see Figure 6A).

3.5.3 *EEG recording.*

We recorded scalp EEG using a 64-channel Biosemi ActiveTwo system (Biosemi, Amsterdam, Netherlands). EEG analog signal was digitized at a 2048 Hz sampling rate. During recording, electrode offset was reduced to between ±50 µV for each individual electrode by softly abrading the underlying scalp with a blunt plastic needle and insulating the electrode tip with saline gel (Sigma Gel, Parker Laboratories, USA).

3.5.4 *Brain Computer Interface.*

Overview.

Our setup comprises one decoding computer and one stimulation computer. The stimulation computer continuously displays the visual stimulation (overlapping face and spiral) placed at the center of the screen oscillating in phase opposition. During the real-time experiment (training, experimental and control phases), the decoding computer continuously receives EEG data and loads them in a buffer for on-line analysis (Oostenveld et al., 2011). Data are also saved for later offline analysis. During on-line analysis, the decoding computer outputs correlation scores for both items presented on the stimulation screen to the stimulation computer. At the end of each trial, the stimulation computer decides based on the correlation scores which stimulus has been preferentially attended during the decision phase

(See Figure 6B).

Decoding procedure.

The decoding model was inspired by backward models of stimulus reconstruction used in recent psychoacoustics studies (O'Sullivan et al., 2015). The decoding computer continuously receives EEG data at a rate of 2048 Hz. Before further processing the recorded data is down-sampled to 256 Hz. The EEG data is next filtered between 1 and 30 Hz with a one-pass Butterworth filter of order 6 and re-referenced to the average signal. Then from a 3 seconds segment R of EEG data, we try to infer a unidimensional signal $\tilde{Y}$ called the reconstructed signal, that represents the visual stimulus most probably attended by the subject during this segment. Along the reconstructed signal we also compute a representation of the 2 concurrent visual stimuli (the oscillation of the face and the oscillations of the spiral) called Y_F and Y_S (Andersen and Müller, 2010). The reconstructed signal $\tilde{Y}$ and the abstract representations Y_F and Y_S of the visual stimuli are vectors with one value per time sample (256 samples/s in our case). The correlations scores c_F and c_S are obtained by correlating $\tilde{Y}$ and Y_F on one hand and $\tilde{Y}$ and Y_S on the other hand.

$$c_S = \text{corr}(\tilde{Y}, Y_S) \tag{1}$$

$$c_F = \text{corr}(\tilde{Y}, Y_F) \tag{2}$$

The stimulation computer thus receives one pair of correlation score (c_F,c_S) every 250ms. Correlation scores were saved for offline analysis on one hand and used to infer the preferentially attended item of the current trial on the other hand. At the end of the trial, we averaged the correlation scores over the last 1.5s of the decision phase (6 correlations scores) for the face and spiral, respectively. The item having the higher averaged scores is designated as preferentially attended by the participants for this trial.

Then the stimulation computer displays the appropriate feedback given the attended item (guessed by the model described beforehand) and the nature of the trial. This feedback corresponds to the decoded decision during the whole training phase and for informative trials of the experimental phase. For deceptive trials of the experimental phase, the opposite item was displayed as feedback.

The reconstructed signal is obtained by applying a linear operation to the EEG data matrix R of dimensions time by channels. The model comprises a series of so-called lags τ_k that account for how the stim-

ulus experienced at time t influences the EEG data at time $t + \tau_k$. More precisely, we have:

$$\tilde{Y}(t) = \sum_{c,k} w_{c,k} R(t + \tau_k, c) \tag{3}$$

where c stands for channel and k for an index of our list of lags and wck are the coefficients that define the backward model.

The model was trained from EEG data collected during the model building phase. Each trial was labelled with the item (face or spiral) the participants were asked to attend to during this trial. The EEG data was preprocessed prior to the training of the model by applying a common average reference (mean EEG is subtracted from all channels) and filtering between 1 and 30 Hz with the same one pass Butterworth filter of order 6 that we use for online decoding procedure.

We find the coefficients wck of the backward model by solving the regression problem:

$$Y = wX \tag{4}$$

where the regressor variable Y contains EEG data from all trials and channels and Y contains the representation of the attended stimulus at each trial (see details in O'Sullivan et al., 2015). To evaluate the accuracy of this model, we performed a cross-validation on the dataset comprising all trials from the model building phase.

3.5.5 *Experimental procedure.*

The experimental protocol was divided in 3 phases: participants first underwent the model building and the training phase before performing the main phase. The main phase consists of 3 identical blocks to which we add a 4th control block for 16 participants.

Model building phase.

For our BCI to decode in real time the preferentially attended item, we first gathered labelled data to train the participant's individual model. At the beginning of each of the 30 trials of this phase, a target was designated by a letter (F for face, S for spiral) overlapping a fixation cross at the center of the screen for 1 second. Participants were asked to preferentially attend the designated items during the whole 5 seconds of the trial.

Training phase.

Participants were then offered to familiarize with the BCI setup for 30 training trials. As in the previous phase, a target was designated at the beginning of each trial. In addition, as soon as participants

continuously attended the same item for 2 seconds, they received feedback for 500 ms consisting of the attended item surrounded by a black square.

Main Phase.

Participants then performed 480 trials presented in 3 successive blocks of 160 trials of 5 seconds. Visual stimulation was continuously displayed on the screen across trials and disappeared only every 8 trials. No target was designated and participants were encouraged to choose their object of attention (i.e., face or spiral) and to change at will across trials. As for the training phase, feedback was provided for 500ms at the end of each trial about the object participants preferentially attended. Crucially, feedback was only informative (reflecting the allegedly attended item) in 75% of the trials (Informative trials). In the remaining 25% of the trials, the other item was displayed instead (Deceptive trials). During one third of the informative trials and during all deceptive trials, participants were asked to report the object they decided to attend, and to perform a confidence judgment about this decision (very unconfident, unconfident, confident, very confident). This distribution of report requests provides an equal amount of report following informative and deceptive trials to analyse. Furthermore, it ensures that receiving a report request does not inform participants on whether the feedback was informative or deceptive.

Control Phase.

Sixteen participants performed 1 block of 160 trials for the control phase. This phase is the same as in the Main phase except that no feedback was provided at the end of the trials. The control block and the 3 blocks of the main phase were presented in random order counterbalanced across participants.

3.5.6 *Data Processing.*

Internal decision evidence (IE). For each trial, we compute a correlate of participants' internal evidence supporting their recent decision. During the decision phase, our BCI continuously outputs correlation scores associated with each object (Figure 1B). These correlation scores reflect how close are the brain signals from being generated by the observation of the face or the spiral respectively. Then, our proxy for internal (decision) evidence (IE) consists in the absolute value of the accumulated difference between the correlation scores associated with each object over the last 3 seconds of the decision phase, and provides an objective measure of how strongly participants preferentially attended one object over the other one.

Accuracy of introspective reports.

We determined for every trial the preferentially attended item the same way we did it online (See Decoding procedure in Brain com-

puter Interface section of the Methods). We operationalized introspective accuracy as being correct when the reported object matches the object decoded by the BCI (e.g., for instance face reported, face decoded), and incorrect otherwise (e.g., face reported, spiral decoded).

Confidence in introspective report.

At the end of the trials, participants reported the confidence they have in their decision along with the decision itself on a 2×4 points scale. Confidence reports were median-split, reports of confidence 1 and 2 were labelled as "Low confidence" trials and reports of 3 and 4 were labelled as "High confidence". For modelling purposes, confidence was coded as a 2-level factor.

Consistency of internal decision evidence.

We computed an index approximating for each trial the ratio of the internal decision evidence strength over its variance. Our consistency index is thus described by the formula:

$$Consistency = \frac{InternalEvidence}{Var(|c_F - c_S|)} \tag{5}$$

where Internal Evidence is described in the previous paragraph and $Var(|c_F - c_S|)$ represents the bootstrapped variance of the correlation difference over the period of accumulation (3 second before the feedback apparition).

3.5.7 *Statistical Analysis.*

Summary statistics were calculated in Matlab (Matlab 2018b, The MathWorks, Inc.). All other statistical tests were calculated in R (Team, 2012). Before applying pairwise comparison, the Shapiro-Wilk method was used to test for the normality of the data. If the normality hypothesis was not rejected, we applied a two-sided paired Student's t-test to our data. Data containing too many empty values or not meeting normality assumptions were analyzed with Wilcoxon rank test. Holm-Bonferroni corrections for multiple comparisons were calculated with R. Cohen d was calculated using the R effsize library (Torchiano, 2016) for approximating effect size.

Both linear mixed effect models and generalized linear mixed effect models were fitted using lmer4 packages (Bates et al., 2007). To operate model reduction we removed non or least significant terms and compared Akaike information criterion (AIC) of more complex and simplified models. Moreover, we operated a Chi-square test to decide whether the more complex model was significantly better at explaining our data. The reported p-values of each fixed effect of linear mixed-effects models and generalized linear mixed-effects models were obtained with this Chi-square test comparing one model with all the possible simplification obtained by removing a single effect. For each model, we detailed the model reduction procedure (see Section B.3).

4

PARTIAL AWARENESS DURING VOLUNTARY ENDOGENOUS DECISION

4.1 SYNOPSIS

People generally feel in control of their free decisions even when they involve equivalent alternatives. Recent studies unraveled a prospective access to neural precursors of intention to act, allowing people to veto their impending decisions. Yet, whether people can also access the content of their ongoing decision remains a debated question. Here we address this question by tracking through a stimulus reconstruction approach the neural signals predicting participants' future free decisions. Participants were asked to either use or avoid to use the content of their early deliberation when making a decision. We showed that participants were unaware of their impending decision as they could they could not depart from them when explicitly asked. Nevertheless, participants report to be conscious of their decision content. We showed that those reports instead correlate with detection of a neural marker of self-initiated decision. Finally we showed that attention allocation during decision execution could retrospectively promote early deliberation content awareness. By suggesting that people are only partially aware of their impending decision, our study provides novel insight on the metacognitive process supervising free choices.

4.2 INTRODUCTION

Facing two equivalent piles of hay, the story says that Buridan's donkey starved to death, unable to choose one. Even facing with a similar situation, people are able to deliberately pick one among equivalent alternatives and generally feel in control of their decisions (Ullmann-Margalit and Morgenbesser, 1977; Haggard, 2017). In the past decades, the cognitive and neural basis of free decisions have brought a lot of interest (Libet et al., 1983; Haggard, 2008). These decisions can be described along three main axis: their content ("what" dimension), their timing ("when" dimension) and their triggering ("whether" dimension) (Brass and Haggard, 2008). From here, a decision is said "free" if at least one of those dimensions is set by the decision maker independently of the environment. Free decisions have been associated with specific fronto-parietal neural circuitry whose activity predicts forthcoming motor choices and abstract intention (Passingham, 1987; Soon et al., 2008; Soon et al., 2013; Passingham, Bengtsson, and Lau, 2010; Bode et al., 2011). Furthermore, these decisions share common principles with decisions based on exogenous factors (Wisniewski, Goschke, and Haynes, 2016). Indeed, environment based choices have

been shown to stem from the accumulation of external sensory evidence (Shadlen and Newsome, 2001; Soon et al., 2008; Bode et al., 2011). Similarly, voluntary internal decisions are thought to result from the accumulation of the random fluctuation of the neural signals (Deco and Romo, 2008; Schurger, Sitt, and Dehaene, 2012; Maoz et al., 2013; Murakami et al., 2014; Furstenberg et al., 2015).

Yet, those mechanistic accounts do not shed any light on the awareness and control people have towards their impending decision (Block, 1995). One one hand, people can cancel an upcoming action up to 200ms before the movement onset (Schultze-Kraft et al., 2016). This veto faculty reveals early conscious access to the "whether" dimension of free decision. At the neural level, this conscious access has been associated to the monitoring of the random fluctuation of neural signal (Parés-Pujolràs et al., 2019; Schultze-Kraft et al., 2020). On the other hand, awareness of an upcoming decision content seems to lag behind the appearance of its neural precursors (Soon et al., 2008; Soon et al., 2013; Bode et al., 2011). Free decision could therefore be triggered unconsciously. In line with this perspective, free decisions have long been considered as a mere retrospective illusion (Libet et al., 1983; Wegner, 2002; Wegner, 2003).

Thus, whether participants can prospectively access and control the content of their decision is still debated. One possibility is that people could be aware of the content of their ongoing decisions. Indeed, participants can change their decision, even when the presented alternatives are equivalent (Ullmann-Margalit and Morgenbesser, 1977; Furstenberg et al., 2015). Yet, since choice alternatives are comparable, option selection could mostly rely on automatic processes (Block, 1995; Kool, Shenhav, and Botvinick, 2017). Thus, participants might remain unaware of the option they are about to pick until retrospectively inferring it (Wegner, 2002; Wegner, 2003; Brass and Haggard, 2007; Shepherd, 2015). To sum up, it remains unclear whether people can be prospectively aware of the deliberation supporting their upcoming decisions.

One way to address this issue, is to confront the information encoded in participants' decision neural precursors with participants' impending decision awareness. We suggest that participants are mostly unconscious of the content of their forthcoming decision. Nonetheless, participant could be aware that a decision is impeding by monitoring the variability of their neural decision precursors (Parés-Pujolràs et al., 2019). Furthermore, as previously observed in partial awareness situations, attention allocation may retrospectively promotes conscious awareness of latent representation of decision content (Sperling, 1960; Block, 2007; Kouider et al., 2010; Sergent et al., 2013).

Here, we relied on a brain-computer interface (BCI) setup to track participants' decision neural precursor before they take a decision. To probe participants' awareness of their impending decision, we adapt a

process dissociation procedure (Jacoby, 1991). Participants had to wait for a cue to appear (pre-cue period) before freely choosing to preferentially attend one out of two overlapping stimuli (post-cue period). In the mean time, their EEG was recorded and a marker of selective attention was continuously measured. After the post-cue period, participants were presented with a feedback showing the results of their decision. Crucially, if participants become aware of choosing an item before the cue appearance, they should adapt their post-cue decision as follows: In half of the trials, participants must stick with their early choice (inclusive trials, green in Figure 9). On the remaining half of the trials, participants were asked to select the other option (exclusive trials, red in Figure 9). Surely, participants can perform inclusion by being merely driven by their unconscious pre-cue deliberation toward one of the two options. However such facilitation effects would impair them while performing an exclusion, which requires a conscious access to the content of the early decision. Although we are not measuring directly the decision neural precursor, we track their direct consequences under the form of participants' selective attention bias toward one of the alternative items. Importantly, our BCI also provides an instrument to act on the environment without any motoric contribution. Indeed, the decision neural precursors we measure are not intermingled with motor preparation signals that could inform participants about their forthcoming decision (Blakemore, Wolpert, and Frith, 2002; Brass, Furstenberg, and Mele, 2019).

4.3 RESULTS

Can people access the content of their forthcoming decisions? To address this question, we first sought to establish whether the measure provided by our BCI could predict the content of participants' future decision. To do so, we computed a measure of participants' internal bias (IB) toward one item before the cue. Based on a stimulus reconstruction approach (O'Sullivan et al. 2015), our BCI continuously outputs correlation scores associated with each object. These correlations scores estimate how close are the brain signals from being generated by preferentially attending the face or the spiral respectively. Then, IB consists in the accumulated difference between the correlation scores associated with each object over the last 2 seconds of the pre-cue period. Thus, IB reflects how strongly the two options were discriminated by the neural precursors of decision (See Methods). Furthermore, we operationalized accuracy for an inclusive trial as being correct if the item decoded as attended in the pre-cue and post-cue period match and incorrect otherwise. On the contrary, exclusive trials were labeled as correct if the pre-cue and post-cue attended items differ and incorrect otherwise.

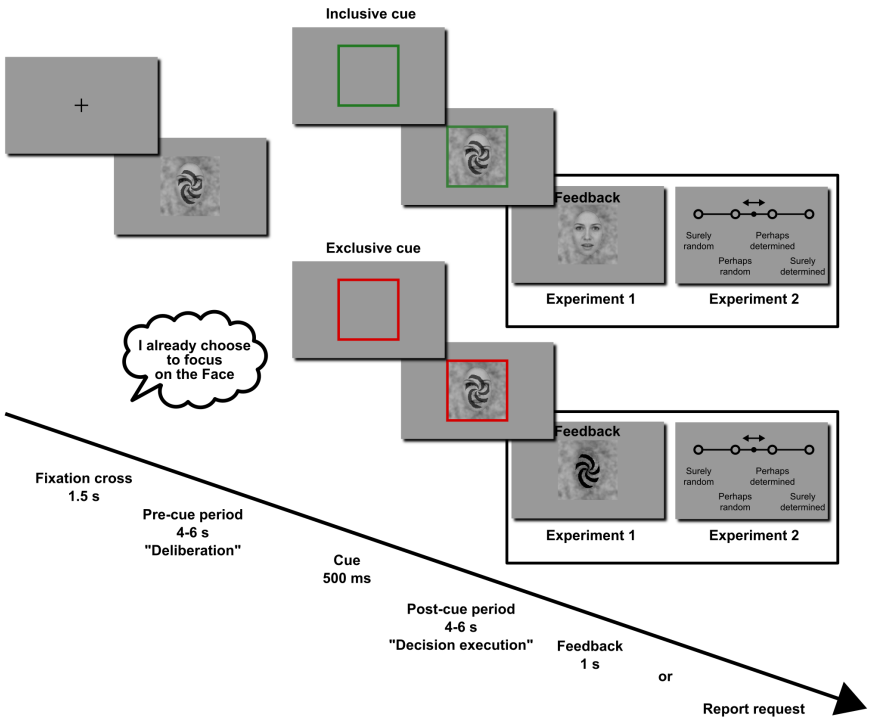

Figure 9: Each trial comprised four phases. 1) Pre-cue period: participants were presented with overlapping face and spiral oscillating at 1.875Hz in temporal phase opposition and asked to wait 4 to 6 s for a cue to appear. 2) Cue: participants were then presented with a colored cue as the visual stimulation disappeared for 500ms and were asked to randomly pick one or the other category by preferentially attending to it. Importantly, if they noticed a premature decision during the pre-cue period, they were asked to keep the item early chosen if the cue was inclusive (green) or the other if the cue was exclusive (red). 3) Post-cue period: participants were then requested to maintain their attention on the chosen category for 4 to 6 second. 4) a) Feedback phase: participants were presented with feedback for 1 s corresponding to the category they preferentially attend during the post-cue period (experiment 1). b) Report phase: participants were asked to report how they choose between the two categories, either randomly or by adapting their choice to the content of their premature decision occurring in the pre-cue period (experiment 2).

Could IB predict post-cue decision? As shown in Figure 10a, IB influences post-cue decision as accuracy increases with IB in inclusion trials (General Linear Mixed Effect Model (GLME): Odd Ratio (OR):1.27, 95% Confidence Interval (CI): [1.08-1.48], χ^2=8.7 , p<0.01). To understand whether the impact of IB results from a conscious processing of early decision content or reflects a mere unconscious facilitation effect, we tested the influence of IB on decision following exclusive cue. Indeed, conscious processing of early decision content would allow participants to correctly revert their impending decision. Conversely, if IB solely act as an unconscious facilitators for future decision, the accuracy would decrease as IB increases. Our results show that despite the presentation of an exclusive cue, participants keep

choosing preferentially the item corresponding to their pre-cue deliberations. Thus, we found an inversion of the relationship between accuracy and IB as underline by the interaction between IB and the nature of the cue (GLME, OR: 0.67, CI:[0.54-0.84],χ^2=12.3 , p<0.001). Indeed, the correlation was negative between accuracy and IB in exclusive trials (GLME, OR: 0.84, CI:[0.71-0.99],χ^2=4.5 , p<0.05). This reveals that although decision derives from early deliberation processes, the latter remain largely unconscious.

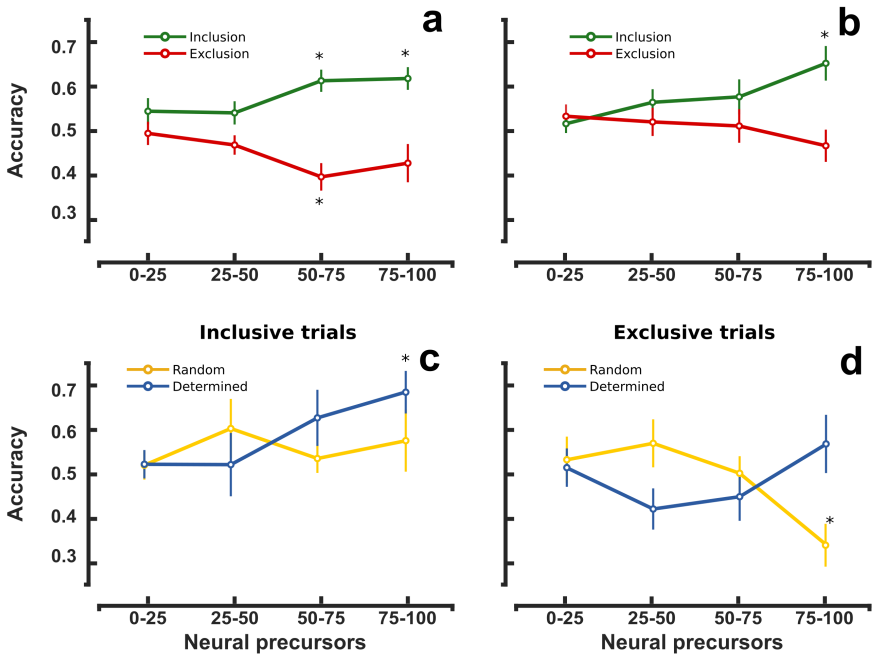

Figure 10: Impact of internal bias on participants' decision awareness.
a-b) Impact of decision neural precursors (IB) on participants' accuracy in inclusive (green) and exclusive (red) trials of experiment 1 (a) or 2 (b). c-d) Impact of decision IB on participants' accuracy in experiment 2 for random (yellow) or determined (blue) choice reports in inclusive (c) or exclusive trials (d). For each participant, we computed the distribution of internal evidence across all trials in terms of percentile. Vertical bars represent standard error to the mean across participants. Stars correspond to intervals of percentile where accuracy significantly departs from chance level tested with paired t-test corrected for multiple comparison by sample permutation (50000 permutations).

In the present paradigm, participants have to adapt their decision depending on whether they noticed early decision process taking place before the cue. Yet, since instructions encouraged participants to wait for the cue before taking their decision, they may have noticed early deliberation process on rare occasion, thereby impairing their decision control faculties. To rule out this possibility, we repeat our first experiment (seeFigure 9) with a probe at the end of each trial. Participants were asked to report whether they pick an item randomly (random choice) or in compliance with our inclusion/ex-

clusion instruction (determined choice). Each possibility came with two level of confidence.

As in experiment 1, pre-cue IB impacts participants' decision although its effect remains largely unconscious (Figure 10b). Indeed, accuracy increases with IB during inclusive trials (OR=1.28, CI=[1.14,1.45],χ^2=16.9, p<0.001) but this correlation was largely overturned in exclusive trials as reflected by the interaction between IB and the nature of the cue (OR=0.69, CI=[0.59-0.82],χ^2=19.3 , p<0.001). Indeed, the correlation between IB and accuracy was negative following exclusive cue (OR=0.89, CI=[0.79-1.00],χ^2=4.2 , p=<0.05). However intriguingly, participants reported an equivalent amount of random (M=125.7, SEM=18.3) and determined choices (M=126.3, SEM=23.7) in total (paired t-test t=-0.02, p=1, Cohen's d =-0.01) and for inclusive and exclusive trials separately (see Appendix C). Those results show that the general poor performances in exclusive trials are not attributable to participants being rarely aware of their early deliberations.

We thus sought to determine whether subjective reports actually capture participants' early deliberation awareness. In both inclusive and exclusive trials, reporting a determined choice increases accuracy (inclusive trials Figure 10c OR=1.16, CI=[0.97-1.38],χ^2=2.6 , p=0.1; exclusive trials Figure 10d OR=1.2, CI=[1.02-1.42],χ^2=5.0 , p<0.05). However, the performance of participants remains below than, or equal to chance level in the exclusive case (random choice: OR= 0.77, CI=[0.65-0.90],χ^2=10.1 , p<0.05; determined choice: OR=1.0, CI=[0.86-1.17],χ^2=0.0 , p=1). On the contrary, performances exceeds chance level in the inclusive case (random choice: OR=1.17, CI=[1.00-1.37],χ^2=3.9 , p<0.05; determined choice: OR=1.45, CI=[1.2-1.76],χ^2=15.4 , p<0.001). Together, those results show that, even when reporting determined choices, participants were not reliably accessing the content of their early decisions.

Nonetheless, should they remain completely blind to their early deliberations, no difference between random and determined reports would have been found. One hypothesis is that participants 's reports reflect a partial access, restricted to certain aspects of their internal variables. More specifically, as participants are known to have prospective access to their intention to act (Schultze-Kraft et al., 2016; Parés-Pujolràs et al., 2019; Schultze-Kraft et al., 2020), we suggest that they can be aware of whether a decision is impending but not of its content.

Active noise reduction has recently been identified as a key neural signature of self-initiated action preparation (Khalighinejad et al., 2018; Khalighinejad et al., 2019). More specifically, noise reduction would be specifically associated with decision process, independently from their motor execution (Khalighinejad et al., 2019). Therefore, we sought to determine whether participant's reports could reflect the variability of their neural signals in the pre-cue period rather than

IB. Since participants' decision are mediated by our BCI, we analyzed the variability of the neural signal driving our BCI. We thus computed the standard deviation of this signal over the pre-cue period of the trial (see Section 4.5).

To determine whether participant's reports reflect detection of an early decision rather than its content, we modeled introspective reports using pre-cue neural variability and IB. We found an absence of effect of IB on introspective reports (Cumulative Link Mixed Models (CLMM), OR=0.98, CI=[0.91-1.05],χ^2=0.3 , p=0.59). Crucially, as the variability of neuronal signals decreases, participants were more likely to report a determined choice (see Figure 11a, CLMM, OR=0.87, CI=[0.78-0.96],χ^2=7.9 , p<0.01). Indeed, pre-cue variance was higher for random (M=-0.03, SEM=0.02) than for determined choice reports (M=-0.12, SEM=0.02), pairwise comparison z=-2.3, d=1.2, p<0.05 see Figure 11b. Thus, participants' awareness reports reflect the variability of the neural signals supporting early deliberation processes. Together, our results demonstrate that despite pretending so, participant are not aware of the content of their early deliberation. Instead, their awareness reports reflect an inflated interpretation of neural noise reduction.

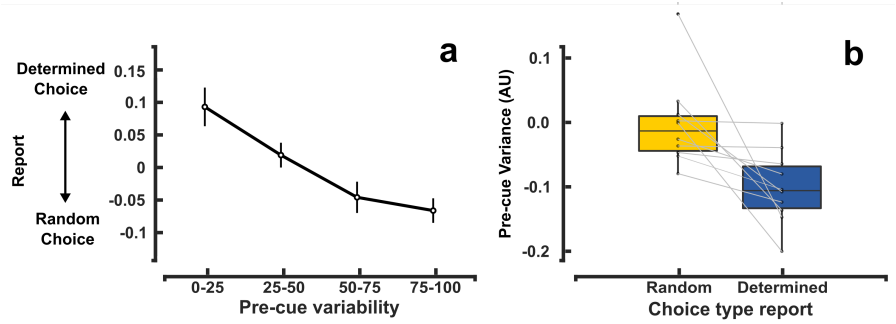

Figure 11: Impact of neural random fluctuations on participants' decision awareness.
a)Impact of the variability of the reconstructed signals on the participants' subjective reports. Reports are given on a 4 points scale where higher scores correspond to determined choice while smaller scores correspond to random choice. Scale was normalized for each participant by referencing each report to the mean. For each participant, we computed the distribution of variability of the pre-cue reconstructed signal across trials in terms of percentile. Vertical bars represent standard error to the mean across participants. b)Level of variability in the pre-cue reconstructed signal for determined and random choice. Star corresponds to pairwise wilcoxon rank test.

In other terms, participants could only access some but not all representational levels of their internal decision processes. Similar situation has been previously documented for external stimulus processing (Kouider and Dupoux, 2004; Kouider et al., 2010), where participants might only reach a partial awareness of their environment

while reporting a richer experience. Importantly, conscious access of the environment in such partial awareness situations can be selectively recovered by retrospective cues (Sperling, 1960; De Gardelle, Sackur, and Kouider, 2009) and attention allocation (Sergent et al., 2013). In line with these studies, we suggest that participants might recover an access to the content of their early decision processes through retrospective allocation of attention during the post-cue period.

Thus, we investigated whether attention allocation can retrospectively promote participants' access to the content of their early deliberations, allowing them to accurately control their subsequent decisions. We computed an index for attention allocation strength (AAS) during the post-cue decision following the same procedure as for IB repeated over the post-cue period. AAS provides a measure of how strongly participants preferentially attended one object over the other one. We model participants' accuracy on exclusive trials in which they reported a determined choice using both IB and AAS. Results show that the impact of decision IB on accuracy is modulated by AAS (see Figure 12b) (GLMER, OR= 1.66, CI= [1.13, 2.45],χ^2=6.7 , p<0.01). Indeed, when AAS is low, accuracy decreases when IB increases (slope (s)= -0.54, 95% Confidence Interval (CI)= [-1.02, -0.06]). However, when AAS is high, accuracy becomes positively correlated with decision IB (s=0.52, CI=[0.04-1.1]) (wilcoxon pairwise test: z=-2.6, p<0.05). These results show that high post-cue attention allocation promotes retrospective access to early deliberation content and allows for the control of the upcoming decision.

Together, our results disclose a discrepancy between the information found in participants' neural signals on one hand and participants' decision awareness on the other. We show that both "what" and "whether" dimension of early decisions were encoded in neural signals preceding decision. Yet, only the fact that a decision is impending was prospectively available to participants' awareness. Strikingly however, the early decision content could be retrospectively access depending on attention allocation.

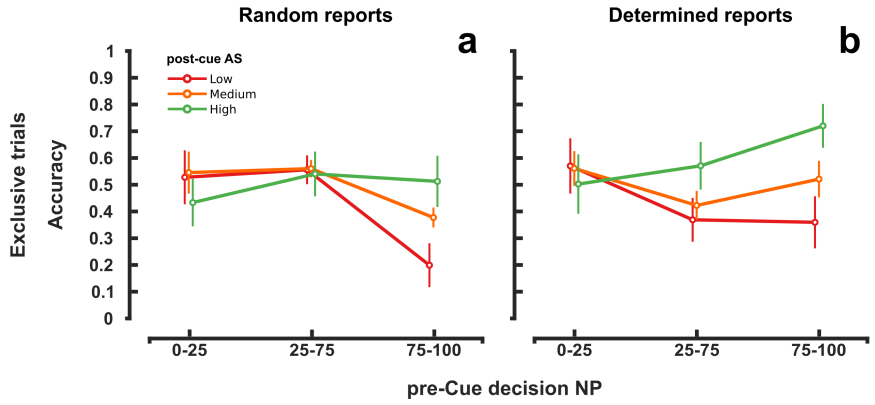

Figure 12: Retrospective attention allocation promote decision awareness. Figures show the impact of decision neural precursor (IB) on accuracy for different levels of post-cue attentional strength (AAs) in exclusive trials. Trials were sorted following subjective report with random choice report (a) and determined choice report (b) respectively. For each participant, we computed the distribution of IB and AAS across all trials in terms of percentile. Low, medium and high levels of post-cue internal evidence correspond to the following percentile intervals: 0-2, 20-80 and 80-100 respectively. Vertical bars represent standard error to the mean across participants.

4.4 DISCUSSION

In the present study, we investigated whether and how participants were conscious of their impending decisions. We compared the information contained in participants' decision neural precursors with participants' control faculties towards their forthcoming choices. We first identify neural signals predicting future decision content. Remarkably, we found that participants had no access to the content of their upcoming decision since they were unable to revert their choices when asked to do so. Intriguingly though, participants often report to have based their decision on their early deliberations content. Yet, reports mostly reflect an inflated interpretation of neural noise reduction rather than an access to the decision content. Nevertheless the content of early deliberation could be retrospectively accessed via retrospective attention allocation. Together, these results show that participants are only partially aware of their ongoing decision.

Our first result is that participants are not able to adapt their decision regarding their preceding internal deliberations. This result extends previous accounts that participants' awareness of their decision to act is largely a retrospective illusion and may not be prospectively accessed and used for subsequent control process (Wegner, 2002; Wegner, 2003; Lau, Rogers, and Passingham, 2007). Yet on the other hand, recent studies emphasized participants' faculty to prospectively access neural precursors of their intention to move, and thereby veto upcoming actions (Schultze-Kraft et al., 2016; Parés-Pujolràs et al.,

2019). Together, these results show that the different aspects of an ongoing free decision are not evenly accessible by participants as the decision unfolds. Indeed, if participants can control whether an action will be triggered, they can't, at least in absence of motor cue, exert a control on what their decision will turn out to be.

Remarkably, participants report to consciously access the content of their early decision despite being unable to properly use it (Block, 1995; Block, 2007). Here, we show these reports correspond to deliberation episode supported by reduced neuronal noise (Brass and Haggard, 2007; Schultze-Kraft et al., 2016; Brass, Furstenberg, and Mele, 2019). This result corroborates recent findings positing that prospective access to an intention to act could be related to conscious monitoring of neural random fluctuations (Parés-Pujolràs et al., 2019). Furthermore, neural noise reduction is thought to be a key precursor of voluntary self-initiated decision (Schurger, Sitt, and Dehaene, 2012; Murakami et al., 2014; Khalighinejad et al., 2018; Brass, Furstenberg, and Mele, 2019). Thereby, similar to motor intention, participants might be aware of whether an early deliberation takes place but remain blind to its content (Schultze-Kraft et al., 2016; Schultze-Kraft et al., 2020; Parés-Pujolràs et al., 2019).

Similar illusory feelings of rich conscious experience is known to happen following the presentation of a degraded perceptual stimulus (O'Regan and Noë, 2001; Sergent and Dehaene, 2004; Block, 2007; Kouider et al., 2010). When briefly presented with a grid of letters, participants reported seeing almost all letters but were only able to report a few (Sperling, 1960; Block, 2007). Such an illusion is thought to reflect a dissociated access between the different representational levels of a perceptual stimuli (Kouider et al., 2010). Here we suggest that this framework could be extended to the awareness of early deliberation processes. Indeed, while participants report a rich experience of their early decision making process, their consciousness is actually limited to the sole knowledge of whether a decision is impending.

Furthermore, retrospective allocation of attention is thought to trigger conscious perception of certain representational levels of external stimulus that would have remained subliminal otherwise (Kouider et al., 2010; Sergent et al., 2013). For the example, retrospective cues can recover reports faculties for a subset of the letter grid evoked above (Sperling, 1960). Similarly participants were able to control the upcoming decision solely when the decision execution was supported by strong attention allocation. According to this finding, we argue that retrospective attention allocation can trigger conscious awareness of certain representational levels (namely the content) of internal deliberation that would have remained unconscious otherwise.

In the present study, we argue that people are generally unaware of the internal deliberation preceding their decisions. Yet, another interpretation of our results could be that people are mostly aware of

the content of their deliberation but fail to select the desired item in the post cue period. Under this view, AAS would rather reflect the success to interact with the BCI than the strength of attention allocation. To account for this alternative hypothesis, we analyzes the effect of AAS on participants' accuracy (see Appendix C) and conclude that correct trials are not attributable to a mere improvement in the use of the BCI setup.

In conclusion, we combine an inclusion-exclusion task with an online decoding approach to demonstrate that participants are only partially aware of their forthcoming decisions. Indeed contrary to their claims, people were only aware of whether a decision was building but not of its content. Our study shed new lights on the mechanism underlying the formation of awareness during the decision process, unraveling a qualitative difference between the information prospectively and retrospectively accessed. Despite the general impression of a rich internal life, participants were only partially aware of the information carried by their decision neural precursors. However, a more complete pictures could be recovered via retrospective attention allocation.

4.5 METHODS

4.5.1 *Participants*

For experiment 1, thirteen healthy participants with normal or corrected-to-normal vision took part in the experiment (4 male; all right-handed; mean age: 27.1 years, SEM = 2.8). One additional participant was tested but was not included in the analysis because technical failures altering the online experiment. Moreover, ten healthy participants with normal or corrected-to-normal vision took part in the experiment 2 (3 male; all right-handed; mean age: 24.1 years, SEM = 3.4). All participants signed a written consent and received financial compensation in exchange for their participation. This experimental protocol was approved by the local ethical committee (Conseil d'évaluation éthique pour les recherches en santé, Paris, France).

Participants performed 288 trials in the main experiment 1. All trials contained a pre-cue, a post cue and feedback phase. and were included in the analysis. In half of the trials the cue requires the participant to perform an inclusion, the remaining half require to perform an exclusion. This led to the analysis of 144 inclusive trials and 144 exclusive trials per participant. Similarly, participants performed 288 trials in the main experiment 2. In the experiment, participants did not receive feedback at the end of each trial. Instead they were required to perform a subjective report about how they made their choice after the cue appearance.

4.5.2 *EEG recording*

We recorded scalp EEG using a 64-channel Biosemi ActiveTwo system (Biosemi, Amsterdam, Netherlands). EEG analog signal was digitized at a 2048 Hz sampling rate. During recording, electrode offset was reduced to between ±50 µV for each individual electrode by softly abrading the underlying scalp with a blunt plastic needle and insulating the electrode tip with saline gel (Sigma Gel, Parker Laboratories, USA). Electrode offset was verified between the different phases of our experiment and between the blocks of the main phase.

4.5.3 *Visual Stimulation*

We used the same visual stimulation that in Chapter 3. we provide in Appendix A a detailed description of its creation procedure.

In the first experimental groups, participants receive feedback corresponding to their post cue selective attention. For 1 seconds, the same stimulation with only the object of attention during the post-cue period was displayed. The word "Result" was also prompt above the visual stimulation.
For the second experimental group, no feedback was displayed at the end of each trial. Instead, participants were asked to report how they made the choice between the two objects at the apparition of the cue. They do so by selecting among four propositions : "My choice was surely random", "My Choice was perhaps random", "My choice was perhaps determined by pre-cue focus", and "My choice was surely determined by pre-cue focus". The choice was performed by moving a dot using left and right arrows of the keyboard, pressing the space-bar validated the chosen position (see also Figure 1).

4.5.4 *Brain Computer Interface*

We use the same BCI that in Chapter 3. For further details about online decoding of attention allocation, see Section 3.5.4.

4.5.5 *Experimental procedure*

The experimental protocol was divided in 3 phases: participants first underwent the model building and the training phase before performing the main phase. The main phase consists of 3 identical blocks.

Model building phase

We first gather labeled data to train the model used in real-time to distinguish current object of participant's attention. To do so, participants were asked to actively bring their attention on one of the two items being simultaneously presented on the screen. Each of the 30 trials lasted 5 seconds, a target item was randomly designated during

1 second by a letter overlaying the fixation cross. The two items were then displayed continuously for 5 seconds during which the participants were asked to maintain their focus of attention.

Training phase

Training phase was meant to both familiarize participants with the BCI setup and with the protocol of the main experiment. Therefore the training phase consists in 20 trials of the main phase of experiment 1. Participants were presented with the same visual stimulation as in the model building phase surrounded by a black square. After 4 to 6 seconds, the black square turned either red or green and participants were instructed to freely choose either the face or the spiral and to focus on it until the end of the trial (4 to 6 seconds after the color change). The color cue remained on screen until the end of the trials. Importantly, if they had chosen the object of their decision before the cue, or if their attention was grabbed by one of the item before the cue, participants were instructed to either keep focusing on this object (inclusion trial yellow cue on Figure 1) or to change the object of their focus (exclusion trials clue cue on Figure 1). To facilitate switch from one item to the other, the visual stimulation disappears for 500 ms after the cue apparition. At the end of the trial, participants received feedback showing the object they attended after the color cue. Feedback was displayed for 1 second with the word "Result" printed above.

Experiment 1 main phase

In this phase, participants performed 288 trials presented in 4 successive blocks of 72 trials of 10.5 seconds. This phase was exactly the same as the training phase. Before beginning this phase, we verify if participants understood the instructions correctly by asking them to summarize the instructions given in the previous phase.

Experiment 2 main phase

For this phase, participants performed 288 trials presented in successive 4 blocks of 72 trials of 10.5 seconds. This phase was the same as the training phase except participants did not receive feedback about their post cue decision. Instead, they were required to perform a metacognitive report about how they made their choice at the cue appearance. They were asked in each trial to choose one among the four following propositions: "My choice was surely random", "My Choice was perhaps random", "My choice was perhaps determined by pre-cue focus", and "My choice was surely determined by pre-cue focus".

4.5.6 *Data Processing*

pre-decision internal decision bias (IB) and attention allocation strength (AS)

For each trial, we compute a proxy for participants' internal evidence supporting their post-cue decision. We first calculate the difference between correlation scores of the two items over the 2 seconds recorded after the cue presentation (between 2.5 and 4.5 second after the cue apparition). We thus compute the accumulated sum over time of this difference. We use the absolute value of the total accumulated evidence to estimate how much internal signals can discriminate between the two items at the time participants execute their decision. This score is referenced as attention allocation strength (AS).

We reiterate this operation on the 2 seconds preceding the cue apparition while participants are required to wait. Although participants were required to refrain from choosing an item during this period, our measure of pre-decision internal decision bias (IB) provides a proxy for how much brain signals were already discriminating between the two items before the cue appearance.

Characterizing the accuracy of a trial

We determine for every trial the preferentially attended item the same way we did online. This ensures our classification corresponds to the feedback the participants receive during the online experiment (See Online analysis in Brain computer Interface section of the Methods). We then determine following the same method the preferentially attended item during the pre-cue period. We then compared the pre-cue and post-cue attended items. An inclusive trial was labeled as correct if the two items match, it was otherwise labeled as an incorrect trial. Conversely, an exclusive trial was labeled as correct if the pre- and post-cue items differ and incorrect otherwise.

Pre-cue neural variability

To assess the variability of the pre-cue internal evidence, we compute for each trial the standard deviation of the reconstructed signal $\tilde{Y}$ (see "decoding procedure" section in Section 3.5.4) over the 2 seconds before the cue.

Analysis of the subjective reports. In the main phase of experiment 2, participants report whether their post-cue decision was random or determined by their pre-cue behavior. Reports were given with two degrees of confidence ("certainly random", "perhaps random"). To investigate the effect of neural signals variability on the participants' reports, the latter were encoded as an ordered factor for cumulative link mixed model analysis, with certainly random choice coded as 1, perhaps random choice coded 2, perhaps determined coded as 3 and certainly determined choice coded as 4. When included as a fixed effect in a general linear mixed model, reports were binarized to reflect

the disjunction between "random choice" versus "determined choice" report.

4.5.7 *Statistical analysis*

Summary statistics were calculated in Matlab (Matlab 2018b, The MathWorks, Inc.). All other statistical tests were calculated in R (Team, 2012).

Before applying pairwise comparison, the Shapiro-Wilk method was used to test for the normality of the data. If the normality hypothesis was not rejected, we applied paired Student's t-test to our data. Data containing too many empty values or not meeting normality assumptions were analyzed with Wilcoxon rank test. Holm-Bonferroni or Tukey correction for multiple comparisons were calculated with R. Cohen d was calculated using R for approximating effect size.

Both linear mixed effect models and generalized linear mixed effect models were fitted using lme4 package (Bates et al., 2007), cumulative link mixed models were computed using ordinal package (Christensen, 2015). Continuous data (pre- and post-cue IE and pre-cue neural variability) were log transformed and centered to avoid convergence issues. To operate model reduction we removed non or least significant terms and compared Akaike information criterion (AIC) of more complex and simplified models. Moreover we operated a Chi-square test to decide whether the more complex model was significantly better at explaining our data. The reported p-values of each fixed effect of linear mixed-effects models and generalized linear mixed-effects models were obtained with this Chi-square test comparing one model with all the possible simplification obtained by removing a single effect. For each model, we detailed the model reduction procedure (see Appendix C).

Post-hoc probing of interaction between continuous variables IB and AS was done using emmeans package (Lenth et al., 2018). We compare the regression slope between accuracy and IB with either low, medium or high level of AS. Levels of AS correspond to the average AS (medium level) plus (high level) or minus (low level) 1.5 standard deviation. Pairwise slope comparison consists of non parametric Wilcoxon test corrected for multiple comparison by Tukey method.

Part III

DISCUSSION

5

GENERAL DISCUSSION

This thesis describes two studies on introspective illusions during free decisions. The central goal of this work was to address whether and how people can become aware of their voluntary, self-driven decisions. More precisely we investigated the conditions under which humans could consciously access the internal variables supporting their decisions. We relied on a novel approach using a Brain Computer Interface to assess participants' hidden decision variables and confront them with their introspective reports. This work provides new insights on the underlying mechanisms of introspection and the emergence of introspective illusion. It also sheds a new light on the relationship between the hierarchical organization of the brain, the measured neural signals and the conscious experience of self-driven decisions.

5.1 SUMMARY OF RESULTS

In Chapter 1, we further detailed our approach to study participants' free decision awareness. In line with recent theoretical models, we proposed that introspection and metacognition can be framed as a Bayesian inference processes on hierarchically organized decision Mele and William, 1992; Pacherie, 2008; Hohwy, 2013. This view of introspection has several consequences.
First, it implies that introspection can not be merely described as an inner sense with privileged access to internal cognitive processes. Instead, in line with the Bayesian brain hypothesis (Knill and Pouget, 2004), introspective content integrates several sources of information from both endogenous processes and the external world.
Secondly, we suggested that the hierarchical organization of decision processes impacts the information available to introspection. We proposed that the different aspects of a free decision could depend on the specific level of the decision hierarchy involved in their computation. This implies that decision-makers could remain unaware of the outcome of their decision when they are deprived of certain levels of the decision hierarchy.

In Chapter 2 we proposed a methodological approach to test these hypotheses by adapting a stimulus reconstruction method to the real-time decoding of top-down endogenous covert attention. We show how our attention-based BCI can isolate specific levels of the decision hierarchy. On the one hand, BCI replaces decision motor execution by providing a unique way to interact with the environment. On the other hand, we limited the contribution of higher order processes by presenting participants with choices between equivalent alternatives. Moreover, our BCI setup allows us to adapt the outcome of a decision

in real-time in order to manipulate the information available during introspection.

In Chapter 3, we investigated the respective contributions of external cues and internal decision variables to the formation of introspective reports. To do so, we induced choice blindness episodes using our BCI setup. We decoded participants' decisions while they performed a simple choice, and provided visual feedback. This feedback was manipulated in 25% of the trials and reflects the desired alternatives in only 75%. Participants were then probed to report their original choice along with their confidence. We managed to elicit Choice Blindness episodes as our participants tend to report the presented feedback as their own choice, even when it contradicts their original preferences. In line with a Bayesian integrative account, we found that the content of introspective reports depend on the reliability of internal variables:

- Unreliable internal variables are dominated by external feedback leading to confabulation reported with high confidence.
- On the contrary when internal evidence supporting the original decision is reliable, participants directly accessed the decision content and easily detected external manipulations.
- Finally, when the reliability of internal variables is intermediate, participants failed to notice external manipulations but their subjective experience was affected as they reported a lower confidence compared to correct introspections.

In Chapter 4, we investigated whether participants can be aware of the content of the deliberation supporting their upcoming free decisions prospectively. Participants were asked to wait for a cue before making a simple two-alternative choice (namely, focusing their attention on one of two items). Crucially, when a choice was made before the cue, participants were asked to either pursue or to revert that choice.

We used our decoding approach to infer participants' preference toward one item before the cue appearance and investigated whether and how these internal biases impact the awareness of their current internal deliberations. We showed that participants could access some, but not all representational levels of their internal decision processes. More specifically, while they could be aware of their impending decision, they could remain unaware and unable to control its content. Our results also show that retrospective attribution of attention could promote conscious access to the content of the initial deliberation.

By addressing the mechanisms of the emergence of decision awareness and its associated illusions, the present thesis provides new insights on diverse subjects. Below, we further detail the implications

of our results for research on the domains of free decision awareness, choice blindness and agentive attention allocation. Yet, the originality of our work also relies on the methodological approach we employed. We will therefore also discuss the methodological challenges encountered when studying complex cognitive processes using BCI. We will complete this discussion by detailing the empirical and theoretical limits of our work, and accompany those considerations by proposing empirical studies to extend our current results and further confront our predictions.

5.2 THEORETICAL IMPLICATIONS

5.2.1 *Implication for awareness of decision*

In the present manuscript, we implicitly considered that dynamic models of intention (Pacherie, 2008; Mele and William, 1992; Brass, Furstenberg, and Mele, 2019) could be interpreted within the predictive coding framework (Friston, 2008; Hohwy, 2013). We suggest that the distal, proximal and motor levels[1] of decision-making represent a manifestation of the Bayesian brain at spatial and temporal scales relevant for decision and action. Conceptually, the distal level is thought to hold the representation of future choices. Those choices are then translated into parameters of the evidence accumulation processes leading to decision (Brass, Furstenberg, and Mele, 2019). From a mechanistic point of view, the distal level modulates the expected precision of bottom-up sensory information and orients the accumulation of evidence toward the desired choice[2] (Friston et al., 2017).

The results presented in Chapter 4 are consistent with this framework. Indeed, when deprived of high order guidance, participants had no access to the content of their decision. Mechanistically, we suggest that the distal level estimates that the precision of evidence supporting both choice alternatives are equal. Naturally during a motor decision, participants could rely on their motor preparation signals to infer the content of their impending decision. This was not the case in our protocol since only the proximal decision level was mobilized. At this level, the brain is thought to adapt long term choices to the immediate context. When all alternatives are equivalent, we suggest that the need for control instances is low. Consequently, the awareness of the decision content would be tenuous. Yet, one might argue that adapting to the immediate context is useful as it might facilitate withdrawal from a dangerous situation. Consistent with this ecological consideration and previous studies (Schultze-Kraft et al., 2016), we suggest that one could become aware that a decision is imminent, allowing her to veto it at the last moment.

1 Referring to the planning, immediate implementation and execution of a decision
2 Gain modulation of superficial pyramidal cells has been proposed to encode this expected precision (FitzGerald et al., 2015).

As a consequence, we suggest that subjects are only partially aware of their proximal decision. The partial awareness hypothesis proposes that representational levels of perceptual processes are hierarchically organized and independently accessed (Kouider et al., 2010). Extending this account to the processing of internal evidence, we propose that different aspects of decision implemented at different levels of the hierarchy are not homogeneously accessed. In other terms, we suggest that before the decision execution, participants can be aware of whether a proximal decision is impending, but can not access its content.

Yet, if participants might not be able to access the neural signals encoding the content of their ongoing decision, we provide evidence in Chapter 3 that those signals participate in post-decision introspective processes. Below, we further discuss how our results provide new perspectives on how internal variables together with various cues could be integrated to form introspective content.

5.2.2 *Implication for interpretation of introspection and Choice Blindness*

In Chapter 1, we proposed that a significant breakthrough in the study of introspection has been to substitute the question of its accuracy by the question of the condition under which a given cue could enter the introspective content. In line with an abundant literature on introspective illusion, choice blindness (CB) paradigms (Johansson et al., 2005) illustrate the impact of exogenous cues on self monitoring processes. In Chapter 3, we went a step further by demonstrating that CB phenomena could result from an integrative process following Bayesian-like principles. Indeed, we showed that unreliable internal decision evidence is dominated by external cues, leading to introspective illusion episodes.

Yet, in the present work, participants were placed in a highly non ecological situation where both distal and motor level of the decision hierarchy were removed while the exogenous cue was under the control of our BCI. As a result, introspection contributions were confined to the proximal level of decision. Participants could extract information about their recent decisions from two sources: the BCI-controlled feedback or their internal attention allocation processes.

Our daily introspective episodes embed higher order belief about both ourselves and the world (Nisbett and Wilson, 1977). Virtually all CB paradigms imply high order priors about the proposed choices, generally under the form of personal preferences or familiarity. Here, we suggest that our Bayesian account could be extended throughout the decision process hierarchy. Broadly speaking, the participation of different cues to introspection could not only be mediated by their intrinsic reliability but also by high order priors (Friston, 2008; Hohwy, 2013). Those priors (e.g. preferences, familiarity etc.) should come in

the form of expectation about the precision of the features processed at the proximal level, thereby modulating their influence on introspection. Therefore, in line with (Hall, Johansson, and Strandberg, 2012; Sauerland et al., 2014; Somerville and McGowan, 2016; Rieznik et al., 2017; Strandberg et al., 2018), a choice involving familiar alternatives would rarely be subject to introspective illusion because the influence of endogenous cues is boosted by distal priors.

Nonetheless, our accounts could be questioned since CB seems unaffected by reward attribution (Hall et al., 2010; Somerville and McGowan, 2016). Although such an approach should inflate the expected precision of the rewarded option, participants are still subject to numerous CB episodes. We proposed that those CB episodes are attributable to social context which can also be translated as a prior on the precision of exogenous cue (i.e. the feedback). A recent unpublished study (Reyes et al., 2018) lends support to this claim by showing that the frequency of CB episodes is proportional to the confidence participants have in the experimental setup. In (Hall et al., 2010), we proposed that the experimental set-up (a supermarket stand) increases participants' trust in the feedback, thus compensating the reward-induced boost of internal variables.

Crucially, as mentioned in Chapter 1, incorrect attributions of precision at a lower level of the hierarchy have been shown to propagate upward, resulting for schizophrenic patients in higher order delusional inferences (Fletcher and Frith, 2009). We suggest that a mechanistically similar phenomenon is taking place in our daily introspective life. Indeed, we shown in Chapter 3 that not only first order content but also second order metacognitive content (i.e., confidence judgments) can be impacted by exogenous cue when the precision of endogenous cue was weak.

Overall,in line with hierarchical Bayesian brain hypothesis (Friston, 2008; Hohwy, 2013; Clark, 2015), introspection results from an integrative process involving first order cue such as internal decision variables and external cues monitoring but also high order priors about the decision maker (i.e. preferences, familiarity etc.) and the external world (i.e. reward value, social factor etc.). Introspective illusions thus reflect the respective expected precision of the cues entering introspective content.

Since our results rely on attention-driven BCI, we must be careful when extending their consequences to ecological decision-making contexts. Indeed, sensori-motor feedback are also integrated during formation of introspective content, further discriminating alternative decisions (Moore and Fletcher, 2012; Legaspi and Toyoizumi, 2019). Yet, by endowing the attention with a direct impact on the environment through our BCI, our work may help to better understand the the process of its agentive allocation.

5.2.3 *Implication for allocation of attention*

In both Chapter 3 and Chapter 4, our BCI substitutes motor action by attention allocation. Attention results, according to predictive coding framework, from the modulation of the expected precision of a given stimulus by the brain. Endogenous decisions, motivations, and other mental actions cast predictions about the precision the different elements of the outer world, thereby reconfiguring the top down attentional landscape (Clark, 2017). Alternatively, change in the precision in the world (e.g. apparition of a salient stimulus) can capture attention through ascending highly precise prediction errors (PE) (Hohwy, 2013). Yet, even in absence of environmental or internal distinction between different stimuli, participants still claim to be able to voluntarily focus on one (Neisser and Becklen, 1975; Ransom, Fazelpour, and Mole, 2017) or switch to other features of their environment (Neisser and Becklen, 1975; Furstenberg et al., 2015).

A first account of such *agentive attention allocation* could be that newly emerging internal beliefs (e.g. being habituated of one stimulus or suddenly wanting to attend the other one) drive voluntary attention allocation through modulation of expecting precision (Clark, 2015; Hohwy, 2013; Clark, 2017). Instead of relying on desires and motivations suddenly emerging, our results suggest that subjects might remain unaware of the target of their voluntary allocation of attention. As suggested in Chapter 4, participants' reports of controlling the target of their attention could stem from an inflated interpretation of neural noise reduction.

At the neuronal level, noise reduction has been identified as a neural correlate of attention allocation (Mitchell, Sundberg, and Reynolds, 2007). Furthermore, attention decreases correlation between neurons encoding a common representation which results in noise reduction in the neural population[3] (Mitchell, Sundberg, and Reynolds, 2009; Cohen and Maunsell, 2009; Pestilli et al., 2011). Such noise reduction could constitute an ideal neural correlate for attention as measured by EEG since the latter averages signals over large groups of neurons.

In Chapter 4, we show that participants' reports of conscious access to their premature decision correlate with a reduced neural noise. Since in our BCI context, decisions correspond to attention allocation, we suggest that participants can consciously access their attention allocation process. Yet, we demonstrate that participants may have not been aware of the object they preferentially attend. In other terms, we suggest that top-down agentive allocation of selective attention could rather be a partially unconscious process retrospectively interpreted

3 Indeed, averaging correlated neurons will not attenuate biases induced by correlated noise. Thus, preliminary decorrelation allows averaging to decrease noise at the population level Serences, 2011

as controlled[4].

In both Chapter 3 and Chapter 4, we considered the external cues to be kept constant. However, attention allocation has been shown to drive change in perception itself (Carrasco, 2011). Yet those changes might be too subtle to participate in attention monitoring processes. Nevertheless, participants' genuine access to the object of their attention was sustained by strong and reliable attention in both Chapter 3 and Chapter 4. We proposed that this result reflects a Bayesian integration process whereby the contribution of a cue to introspection is modulated by its saliency and its reliability. However, an alternative interpretation could be that strong attention allocation drives perceptual modulation that could be retrospectively used to infer the object of attention. Such a hypothesis could be tested by manipulating the contrast of the two items on the screen as a function of the attention that they received.

Overall we suggest that our results are compatible with the predictive coding framework. Moreover, this framework allows us to make specific predictions on the formation of decision awareness and its illusions. In Section 5.4, we will propose some new ideas to address those predictions. Before doing so, we further discuss the benefits and limitations brought by our real-time decoding methodological approach.

5.3 METHODOLOGICAL CONSIDERATION AND LIMITS

5.3.1 *Real time decoding and introspection*

Classical CB paradigm operationalized introspective illusion as the non-detection of a manipulated feedback (i.e. feedback which does not correspond to participants' choices). However, given the impact of external cues on introspection, correct answers following non manipulated trial can also reflect the blind acceptance of external -informative- cue rather than an alleged internal monitoring. In Chapter 3 we tackle this issue by tracking the internal decision variables and confronting them with participants' introspective reports. This approach exploits one of the most interesting aspects of BCI, namely the real time adaptation of experimental conditions to participants' internal state. Yet this novel paradigm brings a certain number of limits inherited from both report (Overgaard and Fazekas, 2016) and no-report (Tsuchiya et al., 2015) approaches and that can be summarized as follows: should we trust participants' report or our BCI outputs?

In Appendix B, we provide control analyses to increase the reliability of our decoding process. We delete all trials suspected of inaccurate classification before repeating our statistical analyses. Yet, since

4 Notice here that we demonstrated that participants were not in control of the object of their selective attention. Yet they might be in control of other aspects of their attention allocation, namely its triggering, its termination or its intensity.

our work involves a novel paradigm, replication of the reported results remains required to confirm them. Furthermore, our approach could have benefited from the measure of a third, independent correlate of participants' free decisions. We further discuss in Section 5.4 how measure of an event related potential (ERP) such as the Feedback Related Negativity (FRN) could provide supplementary information on participants' decision.

Furthermore, subjective reports remain central in the study of consciousness though the optimal way to use and interpret them remains debated (Overgaard, 2015). Indeed identical reports can designate very different subjective experiences across participants but also for one single individual along the experiment. Our analysis copes with inter-individual differences by using a mixed-model approach with the identity of participants systematically set as a random effect. A way to address within subject variations could be to replace our 4 points scale with a continuous scale. Aside from enhancing the precision of the subjective evaluation (Wierzchoń et al., 2019), such scales might provide results that are easier to correlate with our continuous estimation of attention allocation. A second flaw in introspective reports concern the comprehension and good application of the report demands. In Appendix B, we control that participants of Chapter 3 did report their confidence and not other variables such as the strength of their attention allocation.

5.3.2 *Probing complex cognitive mechanisms*

Studying the neural correlates of complex cognitive mechanism can prove challenging for several reasons (Golub et al., 2016):

- We often record only a small portion of the neurons involved in the process of interest.
- The relationship between measured neuronal activity and the final behavior might be non linear.
- Multiple modalities might participate in the control of the process of interest.

BCI appears as a simple solution to tackle these specific issues, especially in the study of motor behavior (Golub et al., 2016). Indeed, classification algorithms consider all the electrodes of the scalp that are engaged in the targeted behavior. Therefore, all the control instances operating on the studied process will eventually affect the decoded signal. Furthermore, this approach defines one single mental action by which participants can interact with the environment. Thereby, we ensure which process participants need to modulate to solve the task. Consequently, aside from the sole classification performance, a crucial question is to know whether our model targets the desired process.

It follows that the dataset on which a participant's model is trained is of primary importance. In the present version of our BCI, partici-

pants were explicitly asked to intentionally attend one specific item during each trial of the model training phase. As a general rule, the outcomes provided by our BCI should be analyzed in regard to the particular (mental) action participants had to perform during model training. Here, we use signals identified in conscious, exogenously driven decisions to study awareness of self-driven free decisions.

As mentioned in Chapter 1, self-driven and exogenously driven decisions may be mechanistically closed (Brass, Furstenberg, and Mele, 2019). However, working with a novel paradigm, we need to demonstrate that our procedure did capture endogenous decisions. We do so in Chapter 3 and Chapter 4 by showing the correlation between our measure and participants' choices. We notably show that participants' neural signals measured by our BCI predict upcoming choices even when participants were unaware of their ongoing deliberation (see exclusion condition in Chapter 4).

Another potential issue in our decoding method is that the cognitive mechanisms captured by our model at a certain time might be process differently during the experiment (e.g. due to fatigue, change in motivation or in strategy etc. Li et al., 2007). Recent investigations on model construction proposed to regularly update participant's model along the task (Shenoy et al., 2006; Li et al., 2007; Huebner et al., 2018). However, such approach requires in general to obtain labeled data with the same concern as during the model training phase: we could target another process (i.e. exogenous driven decision process) that has no reason to have been affected during the main experiment (whose engages self-driven decision process only).

5.4 THEORETICAL LIMITS AND FUTURE DIRECTIONS

In Chapter 1, we suggested that introspection stems from a Bayesian integrative process embedding diverse types of cues. Notably, we suggest that introspective illusions emerge when external deceptive cues dominate the integrative process. As such, evidence supporting our claim could be reinforced. In Chapter 3, we brought indirect proof by showing that illusory introspective content is supported by weak and noisy internal cues. In the following section, we further discuss the influence of exogenous cues reliability on introspection and proposed a supplementary protocol to directly assess their impact.

5.4.1 *Concomitant measure of endogenous and exogenous cues*

Apart from a very few exceptions (Hall et al., 2010), studies using CB paradigm start by inducing confidence in the external feedback by a series of non-manipulated trials (i.e. where feedback correspond to participants' choices) (Johansson et al., 2005; Johansson et al., 2006; Somerville and McGowan, 2016) before introducing a small proportion of manipulated trials. We argue that this procedure increases the

expected precision of feedback, and thereby its weight in the introspection formation process.

Intriguingly, Johansson et al., 2006 rejects the fact that more reliable or precise cues have a stronger impact on the introspective content. Johansson et al., 2006 argues that participants are subject to the same amount of CB episodes whether the underwent the task with a computer or with a human experimenter. Yet, as the argument goes, legerdemain are rare while computer bugs are more frequent. Thus, should the reliability argument stand, participants would detect more easily manipulation in the computer context. However, this argument might be flawed since it ignores the continuous update of precision estimation done at each feedback presentation (Friston et al., 2013). Furthermore, we propose in a follow up experiment to directly address this question.

We claim in Chapter 3 that introspective illusions result from decisions supported by weak and noisy variables. Yet in line with our original proposition, such claim could be extended by suggesting that following a Bayesian hypothesis, both internal decision variables and external feedback impacts are mediated by their respective availability and reliability (Moore and Fletcher, 2012; Knill and Pouget, 2004; Legaspi and Toyoizumi, 2019; Meyniel, Sigman, and Mainen, 2015). However to support such hypothesis, we lacked a measure of availability (i.e. the strength) and of the reliability of the feedback.

We sought to measure the strength of the impact of feedback by assessing at each trial the magnitude of the Feedback Related Negativity (FRN), an Event Related Potential triggered by surprising outcomes (Oliveira, McDonald, and Goodman, 2007; Hauser et al., 2014). Measured over the mid-central area, FRN is thought to reflect the error of prediction associated with an outcome (Talmi et al., 2012; Talmi, Atkinson, and El-Deredy, 2013). Yet, our efforts to extract ERP from the EEG recording remained vain.

Indeed, the emphasis put on attention decoding accuracy led us to opt for a trial structure unadapted to the analysis of post decision ERP. Because participants were asked to minimize eye movement and blink during the presentation of the stimuli, blinks frequently occur just after feedback apparition, corrupting specifically the time window of interest for FRN analysis. Moreover, the feedback presentation was short and embedded in visual noise, limiting the maximal amplitude of potential ERP. Finally, the feedback was revealed by continuing the animated sequence with only one item. Therefore, the feedback might not reach its maximum visibility at the same time for every trial, blurring the result of our ERP approach.

Here, to control the reliability of the feedback[5], we propose an experiment where the percentage of deceptive feedback varies across blocks. We propose a protocol with 3 blocks presented in randomized order with correct feedback in 75%, 50% and 25% of the trials respectively (see Figure 13A).

We aim first at reproducing our results in the 75% block. In the block of lower reliability, we predict that deceptive feedback will have a smaller impact on introspective reports. We first expect an interaction between the nature and the reliability of the feedback when modeling introspective reports accuracy. Indeed, deceptive feedback should impair more strongly the introspective reports in the 75% block than in the other [6]. Metacognitive failures observed in our original experiment should also be impacted by modulation of feedback reliability. Indeed, we suggest that confidence attributed to decisions supported by weak and noisy internal variables mostly reflect the confidence attributed to the external feedback. Therefore, we predict that the negative correlation between confidence and internal variable reliability would be flatten while reliability of the feedback decreases (see Figure 13C).

5.4.2 *Probing the limits of awareness of decision content*

In Chapter 4, we investigate how participants can be conscious of the content of their impending decision. We operationalized participants' decision awareness through their ability to revert their choices when asked to. Yet, this approach relies on the hypothesis that consciousness and control faculties are indissociable (Block, 1995). Yet, in other theoretical paradigm such as higher-order thoughts theories, control can be operated while the relevant cognitive process remains unconscious (Carruthers, 2011). This notably account for strategic behavior adopted by hypnotized participants (Dienes and Perner, 2007). Therefore, a replication of Chapter 4 results would benefit from a direct probing of the content of the decision.

5 From a predictive coding perspective, modulating the reliability of the feedback will influence the hyperpriors about the expected precision of external outcome. It would thus modulate the participation of those external cues in the process of model updating to account for incoming PE. In other terms, a feedback associated with low expected precision will not have much influence on our representation of the world.

6 As shown in Figure 13B , the positive impact of informative feedback should also be reduced, resulting in accuracy decreasing with feedback reliability in the informative condition.

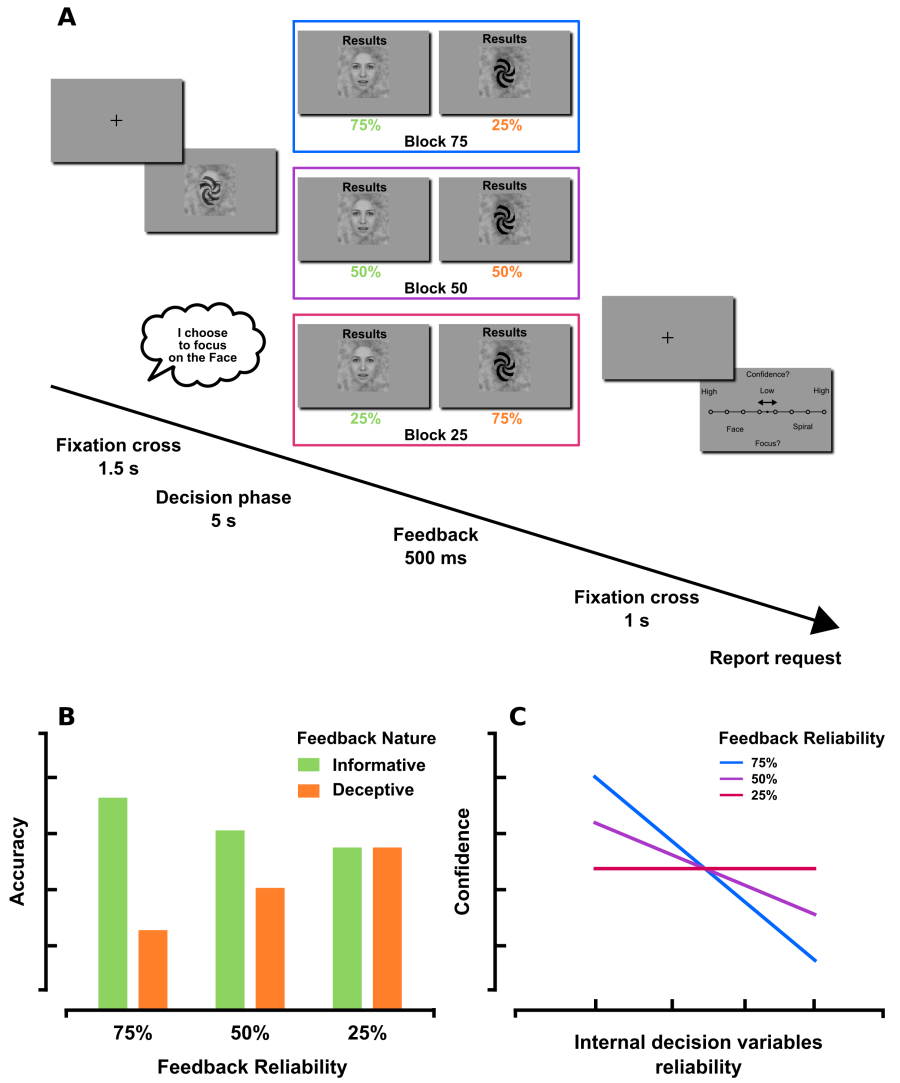

Figure 13: Modulating Feedback reliability in BCI-induced Choice blindness paradigm
A) Experimental paradigm. Each trial consists in the three same phase as in Figure 6 with extra fixation time between feedback and report phase for ERP analysis purpose. Moreover, the feedback now consist in fixed images of maximal visibility instead of the animated sequence. B) Prediction on participants' introspective accuracy. Accuracy is computed as the proportion of trials in which explicit reports correspond to the decoded decision. General mixed effect modeling is used to assess the effect of feedback reliability and feedback nature on accuracy. C) Prediction on participants' confidence during introspective illusion. Confidence is evaluated on a 4 point scale. Analysis is restrained to incorrect reports following deceptive trials. Internal evidence variability could either be continuous or discrete (by median split).

Noteworthy in the second experiment of Chapter 4, we probe participants on whether their decisions complied with the inclusion/exclusion demand. This formulation is questionable as it encloses two sub-questions. First did the participants notice a premature decision, second did the participants manage to select the desired target ac-

cording to the task demand. Tying the two questions together might have created a confusion for participants. Here we propose a protocol to further probe the awareness of conscious content.

Our task consists in an implicit process dissociation procedure protocol where the decision is imposed by the BCI. Participants are asked to wait for a cue to appear at a random time before choosing to preferentially attend one among two items. During the waiting time, our BCI track participant's early deliberation process. On 30% of the trial, a neutral cue indicates that participants can freely choose to attend one item (free cue). On 30% of the trials, participants are ask to attend the item corresponding to their early deliberation (inclusive cue). On 30% of trials, participants were asked to focus on the item opposing their early deliberation (exclusive cue). Participants have to focus on the target designated by the cue until a green cross appears. Then participants are asked to report what object they intended to choose when the cue appears along with their confidence. Finally in the last 10% of trials, participants were asked to immediately report the object of their current deliberation along with confidence (probe cue) (see Figure 14A).

We predict that both the nature of the cue and internal variables would impact participants' reports at the trial end (Figure 14B). Furthermore, we suggest that the influence of internal variables will be mediated by post-cue attention allocation (Figure 14C). Finally, we hypothesize that if early action selection could impact introspective report (Chambon and Haggard, 2012), it could hardly be consciously accesses as decisions unfold. Comparing reports following free cue and probe cue, we hypothesize that impact of internal variable will be reduced in the immediate report condition (Figure 14D).

5.4.3 *Neural noise and awareness*

In Chapter 4, we first suggested that participants could be aware of whether a decision was impending. However, we only brought indirect evidence in support of such a claim. Indeed, our argument relies on a correlation between participants' report and an identified neural precursor of voluntary decision (Khalighinejad et al., 2018; Khalighinejad et al., 2019). As mentioned in Chapter 4, noise reduction monitoring has already been suggested to directly impact awareness of forthcoming motor decision (Parés-Pujolràs et al., 2019; Schultze-Kraft et al., 2020; Schultze-Kraft et al., 2016).

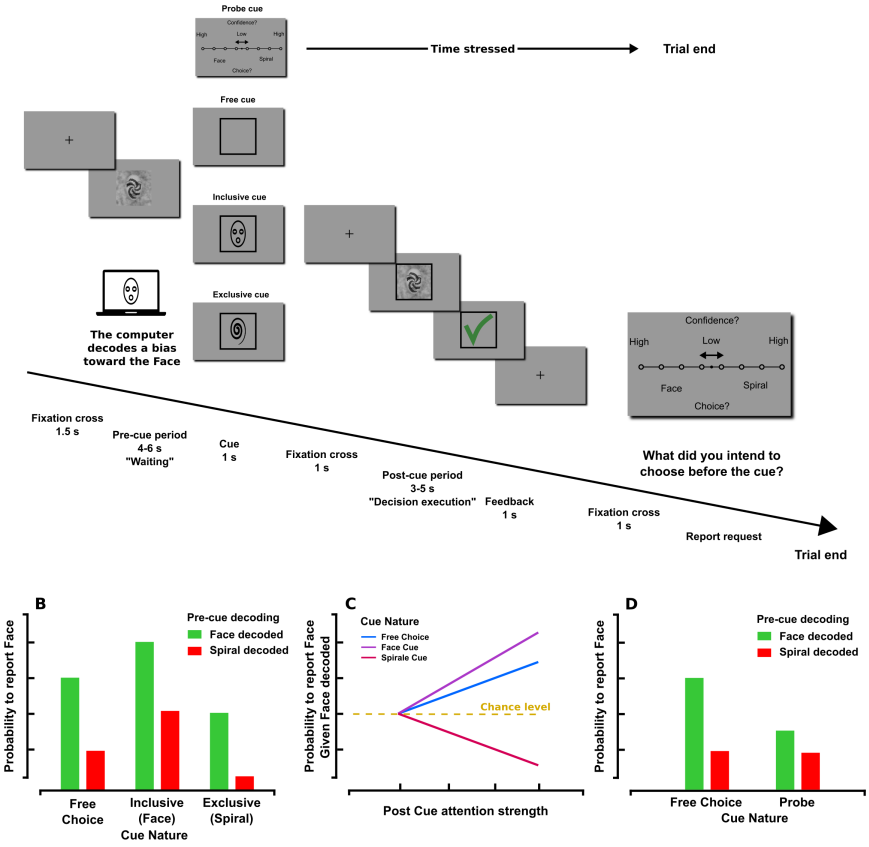

Figure 14: Probing awareness of decision content along the decision process A) Experimental paradigm. Participants wait for a cue appearing at random time to pick one item on the screen while the BCI tracks their ongoing preferences. They are then asked to choose the item either comforting (inclusive cue) or opposing (exclusive cue) their original preferences or to pick one freely (free cue). A green validation sign appears after participants focus enough on their target. Participants are finally probed to report their pre-cue preferences. On an extra 10% of trials, probing was done at the cue time (probe cue). B,C,D) Predictions. We show an example where a "face choice" has been decoded in the pre-cue period. B)Influence of internal and external cue on report. C) Influence of post-cue attention of decision content awareness. Confidence is evaluated on a 4 point scale. Analysis is restrained to incorrect reports following deceptive trials. D) Emergence of decision content awareness. We compare preferences reports at the time of the probe and after the pre-cue period to assess the effect of decision execution on decision awareness.

Yet this claim rests upon an indirect three steps reasoning. First, a competence model proposed that RP reflects the accumulation of neural random fluctuations[7] (Schurger, Sitt, and Dehaene, 2012). Second, a correlation between RP detection and participants' intention awareness was established. Third the authors concluded that participants could have some access to their neural random fluctuations.

7 This account has been criticized notably because it predict that RP-like signals should be widely present, which does not appear to be the case (Travers et al., 2020)

Here we show a direct correlation between neural noise reduction and modulation of decision awareness in a non motor paradigm. Yet, as mentioned in the previous section, our probe did not target specifically the awareness of an upcoming decision. Further investigations are thus required to directly assess the link between random neural fluctuations and decision awareness

5.4.4 *Toward a metacognitive prosthesis*

We show in Chapter 3 that introspective illusions could be accompanied by metacognitive failures. In other words, subjects might feel confident for their illusions. Such dissociation between performance and confidence have already been documented and are suspected to result from a coupled yet distinct sampling of evidence for first order and metacognitive reports (Lau and Passingham, 2006; Wilimzig et al., 2008; Graziano and Sigman, 2009; Rahnev et al., 2011; Bona and Silvanto, 2014; Vlassova, Donkin, and Pearson, 2014).

Metacognition in a predictive coding framework is thought to control the precision accorded to the different cues participating in our model of the world. From an evolutionary perspective, according a large credit to external cues when making inference on your own behavior could save cognitive resources with only a limited chance of error (Shenhav et al., 2017; Kool, Shenhav, and Botvinick, 2017). Indeed, experimental environments aside, external cues are generally informative about our recent decisions. Yet, the emergence and spreading of new technologies such as BCI can be a rule changer by placing cognition in novel and highly non ecological situations. As we have shown, participants might be unable to monitor some aspects of their mental actions, thereby impairing subsequent behavioral control. In such context, the advantageous posture of a metacognition according credit to external cues could be revised.

Here, we proposed that BCI related metacognition should be trained in order to increase the precision of internal cues. Metacognition plays an important role in the development of expert skills (MacIntyre et al., 2014). We describe a protocol for a metacognitive prosthesis whose function is to increase the users' control over their BCI actions. Previous studies have proposed to train participants to guess the results of their BCI actions in the absence of feedback with only limited success (Schurger et al., 2017). Here, we propose to improve metacognitive faculty by rewarding participants depending on the precision of their internal variables.

Participants would first undergo a probing session which consists in a block of the main experiment of Chapter 3. At the end of this session, we would compute the distribution of the precision of internal

variables during the decision phase[8].
Participants would then be presented with a training session. They would be asked to optimize their gains by freely choosing one among two items before receiving a reward. Yet, unknown to the participants, the reward would not be associated with the choice but rather increases with the precision of internal decision variables (see Figure 15). To continuously adapt the setup to participants' performance, the distribution of the precision would be dynamically updated during the task .
After the training session, we would repeat the probing session. We predict that the number of choice blindness episodes should decrease after the training sessions, in proportion with the increase in internal signals precision. To control for habituation, fatigue and other adaptation phenomena, a group control is presented with a third probe session instead of the training session.

5.5 FINAL CONCLUSION

We usually consider that we have privileged and exhaustive access to our internal mental life. In particular, we feel in control of our voluntary action and provide extensive justification for them if needed. Yet, illusion and external influences are known to be integral parts of our self-knowledge processes. Therefore, investigating whether and how participants can access the neural underpinnings of their decisions is a challenging issue. In the present thesis, we opt for a novel strategy to address this problem. We start from a conception of decision processes embedded in a hierarchical Bayesian theory of the brain. Then, we design a non ecological decisional paradigm where participants make simple free choices between equivalent alternatives and where motor actions were replaced by attention allocation process. This approach allows us to dissect both the origin and the dynamic of the formation of decision awareness. We suggest that awareness of our recent decision involves endogenous and exogenous cues integrated following Bayesian principles. Furthermore, we suggest that conscious control of decision is distributed across the hierarchical levels of the decision organization. Our work provides new empirical insights on the formation of the decision related metacognitive content. More generally, this work emphasizes that real-time decoding and BCI based experimental work can be used on a regular basis to investigate complex cognitive processes.

8 Precision of internal cue is operationalized as the standard deviation of the reconstructed signals computed along the decision phase.

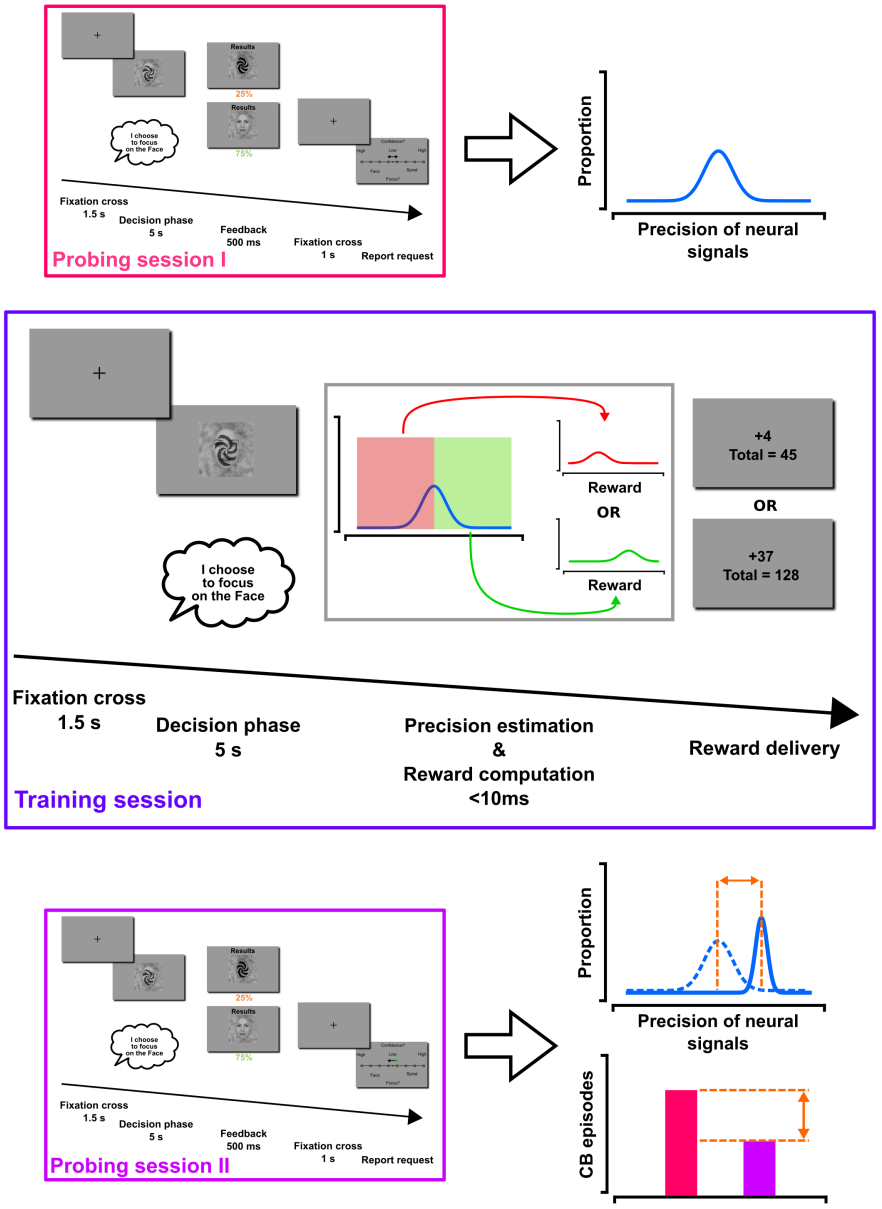

Figure 15: Metacognitive prosthesis

Top) First probing session. See Chapter 3 for further details. Precision of internal cues is estimated by computing the inverse of the variance of the reconstructed signal over the 5 s of the decision phase (see Section 4.5 in Chapter 4 for further detail. Middle) Training session. Participants are asked to maximize their gain by freely picking one over two items. Reward is estimated based on the precision of internal evidence. If the precision of internal cues supporting the current decision falls below the mean on the distribution (red zone), reward is drawn from a distribution centered around a small value (red reward distribution). On the contrary reward is picked from a (green) high mean distribution if precision of current trial internal cue are high (green zone) Down) Second probing session, identical to the first one. The number of CB episodes are compared between the first (light purple) and the second (dark purple) probing session. We predict a correlation between lower CB episode number and an increase in the precision of neural signals (orange arrows).

Part IV

APPENDIX

A

METHODOLOGICAL FOREWORD: THE VISUAL STIMULATION

A word here on stimulus reconstruction going online

We present here the process we followed to create our visual stimulation. To begin with, we will expose the specifications that our visual stimulation must meet. These specifications respond at the same time to the technical constraints imposed by the ssVEP mediated BCI and to the experimental requirement that our research imposed. We will then detail the step-by-step procedure of creation of the visual stimulus.

A.1 SPECIFICATIONS

Technical specifications define the requirement our stimulation must meet to elicit brain signals that can be decoded by our BCI. On the other side, the research-related specifications are shaped by the theoretical question we aim to assess in our experimental work. As we have seen in Section 2.1 of Chapter 2 , diverse form of selective attention can modulate ssVEP and thereby serve as brain input in the BCI loop. In all our experiment, participants were invited to make a free decision of choosing to selectively attend on of the two proposed alternative on the screen. To keep control on the variables participating to metacognitive process, we choose to rely only on features-based selective attention. Thus, to control the information participants receive about their decision, the visual stimulation should limit other form of selective attention to be involved in the decision process. Noteworthy, our work is guided by the scientific question and not by the optimization of our BCI performance. Therefore the research-related specifications prevail in the stimulation creation process. Nonetheless, as a reliable decoding is a key condition to address our scientific question, performance of the BCI remains a crucial factor. Thereby, we can compile a list of requirement for our stimulus:

- Experimental specifications
    1. Our stimulation needs to contains two interactive elements.

       → Our stimulation presents two alternative choices: a face and a spiral.

    2. To avoid spatial attention to inform participants about their current decision, the two elements have to occupy the same spatial location.
       → We present the two elements overlapped. Moreover, the

salient part of each element were located at the center of the display which correspond to fovea of participants.

3. Feature-based attention should be easy to maintain for long period on a single element(10-20 seconds).
   →Item were animated in short looping sequence. This animation help sustaining feature-based selective attention for long period.

4. Participants should be able to change their decision at will. Therefore the attentional cost for switching from on item to the other should remain low.
   → We choose items that were unlikely to be found overlapping at a similar scale in natural scene. Therefore, the hyper-prior associated with seeing the mixture of both elements was low (Ransom, Fazelpour, and Mole, 2017; Clark, 2017). One can thus switch easily from one item to the other without having its attention grab by their superposition (Neisser and Becklen, 1975).

5. The two elements should be present at the screen an equal amount of time to avoid introducing a bias in the decision process.
   → Face and spiral oscillate at the same frequency. Therefore they appear an equal number of frame on the screen.

- Technical specification
  1. Oscillation of the elements must elicit distinguishable visual evoked potential.
     → Although the modulation frequency of the both elements was identical, they were oscillating in temporal phase opposition. Thereby, both item produce distinguishable steady state visual evoked potential.

  2. Oscillation have to be visually comfortable to allow for 2 hours long experimental sessions.
     → To obtain smooth blending of our features, we used a method called sweep-ssVEP (Regan, 1973; Ales et al., 2012) that we further detail in Section A.2.

A.2 CREATION PROCEDURE

THE SWEEP-SSVEP APPROACH The sweep-ssVEP approach (Regan, 1973) consist in eliciting ssVEP by systematically varying the visibility of a stimulus. For this variation to be smooth while eliciting strong and reliable ssVEP signals, visibility is controlled trough spatial phase scrambling Ales et al., 2012. This method keep power spectrum fixed for all images and vary only the spatial phase from

a random phase (minimal visibility) to the original item phase (maximal visibility). This manipulation preserves distribution of low-level image statistics like the mean luminance between the different level of visibility. As an example, the spatial phase of the spiral image will be cyclically scrambled and descrambled to produce ssVEP at the requested frequency (see Figure 17).

A.2.1 *Background (see Figure 16)*

We first created a background matching the overlapping face and spiral in average power spectrum and luminance. This procedure minimize abrupt transition between background and target elements and produce the impression that face and spiral are smoothly emerging from the background. We extracted 12 images from an animation representing a face opening and closing the mouth and 8 images from an animation representing a rotating spiral. We then cropped each image and smoothed the contour by applying a gaussian filter. Next, we computed the average power spectrum separately for the set of faces and spirals images (1 in Figure 16). We then create one image of background for each set by applying inverse Fourier transform to the average power spectrum with a random spatial phase (2 in Figure 16). Finally, we superimpose both face and spiral-extracted background in one single image (3 in Figure 16). As explained below, we generated a total of 96 frames of visual stimulation that were played on a loop. Therefore we produce 96 images for stimulus background by drawing different random phase for each images.

A.2.2 *Animation and phase modulation (see Figure 17)*

To compute a sequence of animation that can be play in loop, both animations should start at the beginning of the sequence and end simultaneously. To do so we create a sequence of 8 repetitions of the face animation and 12 repetition of the spiral animation, resulting in a 96 frames long animation. Following this, for each frame, we select the face and spiral image corresponding to this frame animation step.

To elicit visual evoked potential, we then determined the degree of phase scrambling respectively for face and spiral at each frame of the animated sequence. We first computed 20 uniformly distributed level of phase scrambling for each image. The level of phase scrambling that will actually be used for a given participant are determined on the next step . This distribution was obtained by linearly interpolating phase angle and choosing the shortest distance between the phases. The interested reader will find details and motives for this methods in (Ales et al., 2012).

Animated streams are easier to attend if they never completely disappear. On the other hand, the larger the range of used phase, the stronger will be the evoked signals. Therefore, for each participant we found a trade-off between the strength of the evoked signals and the

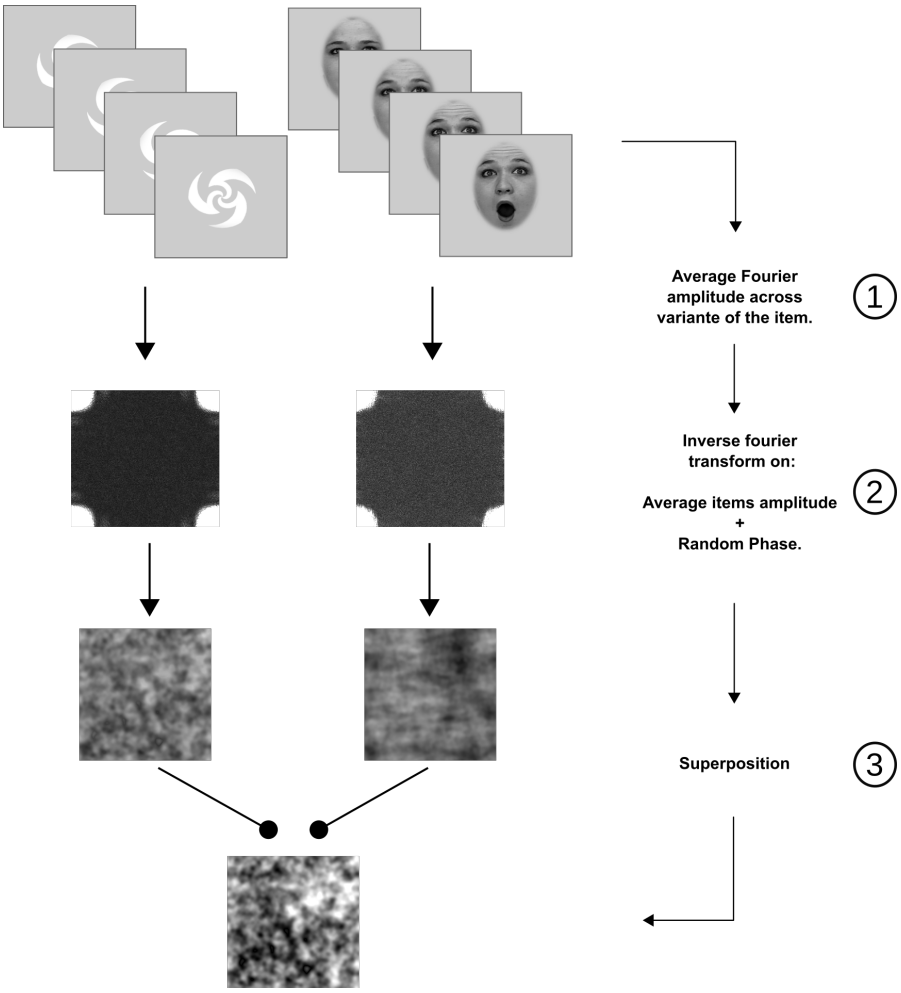

Figure 16: Creation of one frame of visual background.

easiness to attend both streams. This procedure determined the minimum and maximum level of scrambling between which the phase of each elements will vary during the animation. We then compute the precise temporal sequence of phase variation for each item separately. We use a number of frame for one cycle of variation that divide 96 and choose 32. Therefore, both item elicit ssVEP at the frequency of 1.875 Hz and in temporal phase opposition.

As switching from one stream to the other should be effortless, we keep the duty cycle of the phase modulation low (at 0.2). Thereby, the most visible version of each elements are not displayed for more than two successive frames, avoiding unintentional exogenous capture of the attention by one element.
After this step, we end up with the step of the animation and the level of phase scrambling to apply respectively for face and spiral for each of the 96 frames of the visual stimulation(see Figure 17).

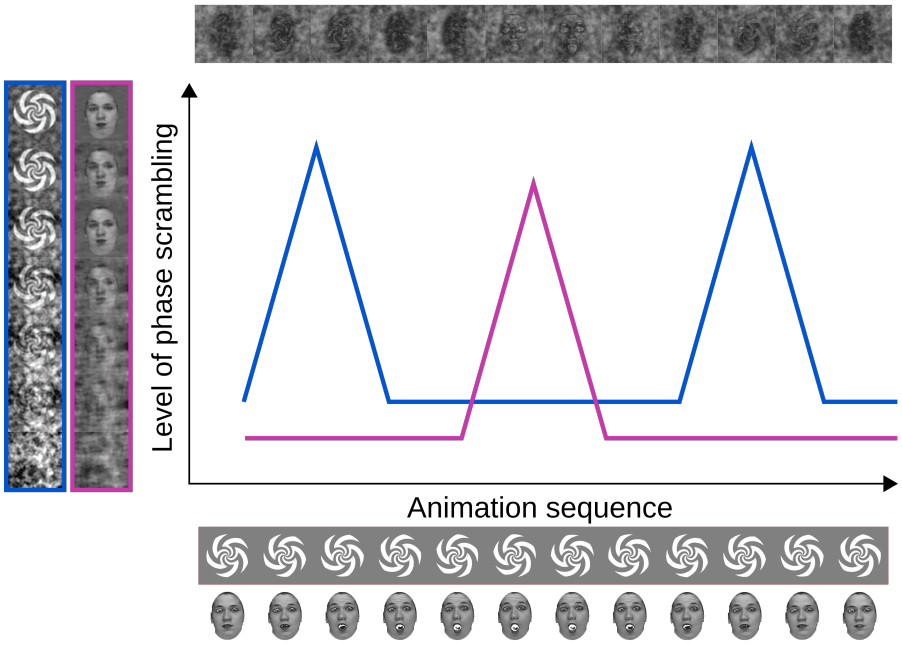

Figure 17: Animation and phase modulation.

For each frame, show the level of phase scrambling and the steps of animation applied to face and spiral respectively to create the image. The process is describe in detailed in Section A.2.2. Y axis: level of spatial phase scrambling. X axis: steps to obtain the animated stimulation. Blue and violet curves: phase scrambling modulation of the spiral and face respectively across time. Below the graphic: successive steps of the animation. Left to the graphic: levels of phase scrambling for our two stimulus. Above the graphic: resulting stimulation.

A.2.3 *Single frame creation (see Figure 18)*

For one given frame we begin by the same steps as in background creation: we obtained average power spectrum over the entire set of face and spiral images separately (1 in Figure 18). We then select for the face and the spiral the image corresponding to the animation step at this frame. The we apply the following procedure on the face and the spiral separately (example is given for the spiral):
We first compute an image having the phase of the original spiral and the power spectrum equal to the computed average over all spirals of the animation (2 in Figure 18). We then apply spatial band-pass filter to remove the highest spatial frequency (3 in Figure 18). Indeed previous studies shown that strength of evoked signal peak for relatively low spatial frequency (Arakawa et al., 1999; Zhu et al., 2010; Norcia et al., 2015). To ensure that both elements occupy the same spatial position, we apply the same alpha mask to each of them along with the opposite alpha mask to the background selected for this frame (4 in Figure 18). Finally, the two separate elements and the background were concatenate to form one of the 96 frames of our visual stimulation (5 in Figure 18).

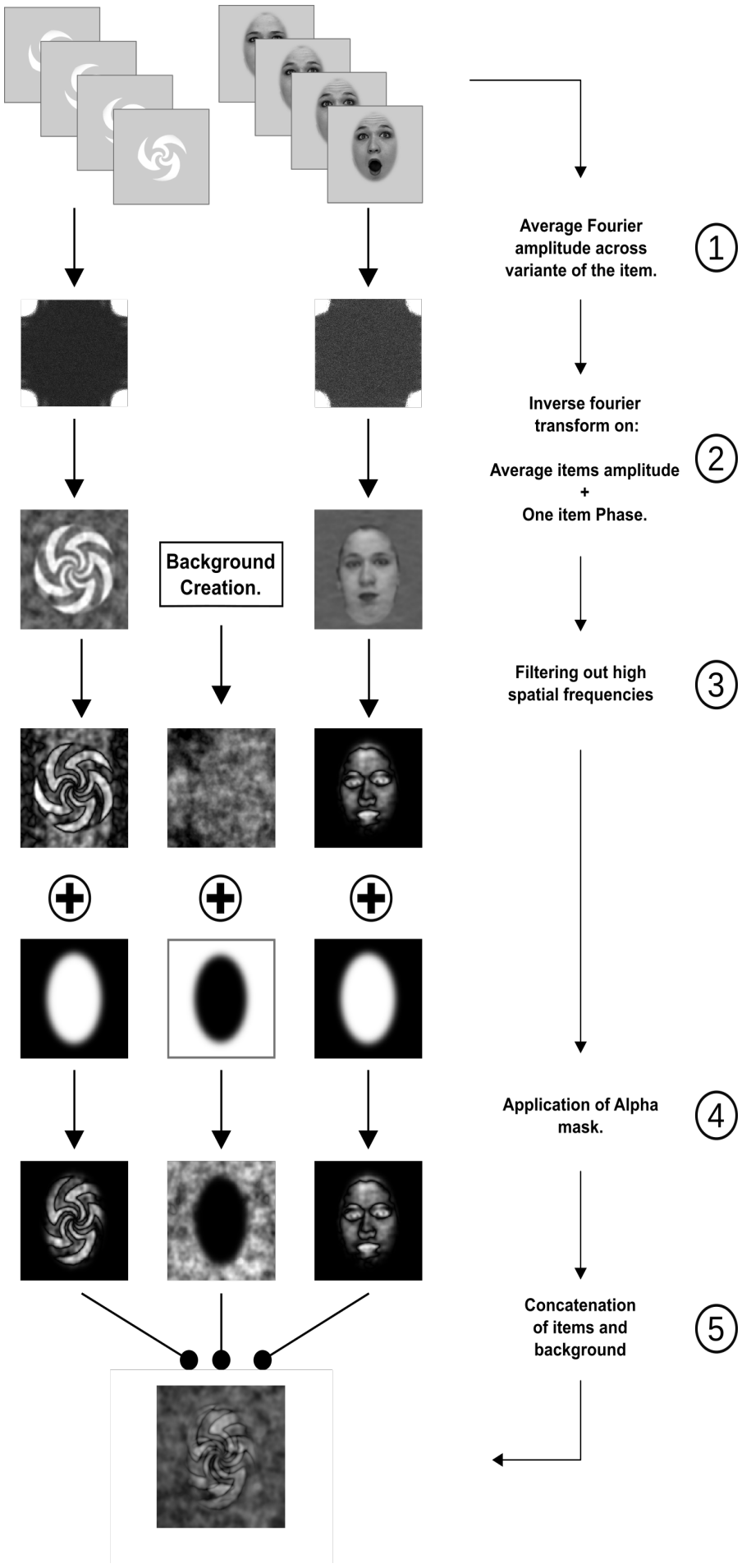

Figure 18: Creation of one frame of visual stimulation.

B

EXTENDED DATA FOR CHAPTER 3

B.1 RESULTS

B.1.1 *Effect of internal decision evidence on introspective accuracy in the absence of feedback.*

As exposed in the main text, we confirmed that internal decision evidence has an influence on introspective accuracy independently of external cues by running a control experiment where participants were not presented with the outcome of their decision at the end of the trials. As shown in Figure 19A, positive correlation between accuracy and IE was not due to the presentation of a feedback as it remains present without feedback (GLME, OR=1.19, CI= [1.13-1.26], χ^2 = 111.1, p<0.0001, Table 2 in Section B.3). Interaction between the feedback type and IE was not significant (GLME, χ^2 = 2.2, p>0,3), confirming that feedback cues modulate the accuracy of introspective reports regardless of the internal information available during the decision phase (see Figure 19B).

B.1.2 *Distinguishing accurate introspection from confabulation.*

To better understand the underlying mechanisms of overconfident confabulations, we first needed to identified trials in which confabulations could be distinguished from accurate internal monitoring. We remarked that following a deceptive feedback, a confabulatory report of the displayed feedback instead of the original decision is labelled as an erroneous trial, while a correct introspection results in a correct trial. Alternatively, following an informative feedback presentation, a confabulatory report of the displayed feedback (i.e. a report unrelated to introspection of internal decision variable) will be labelled as a correct trial as feedback correspond to the original decision. Therefore, confabulations should be mixed up with correct trials when occurring after the administration of informative feedback cues. This remark was confirmed by a significantly different relationship between consistency, accuracy and confidence for deceptive and informative feedback (linear mixed effect model (LME): Estimate=-0,23, CI=[-0,39 -0,06], χ^2 =6.9, , p=0,009, Section B.3 Table 4). Thus, to investigate the relationship between confidence and internal decision evidence consistency during confabulated and correct reports, we regressed consistency against accuracy and confidence for deceptive trials (see Chapter 3).

B.1.3 *Relationship between confidence and consistency of internal decision evidence following informative feedback cues.*

According to our hypothesis, confabulations occur when internal evidence supporting the decision is weak and noisy. However, consistency should not vary with accuracy nor confidence following informative feedback administration, because trials labelled as correct regroup both accurate introspection and reports driven by external cues. As expected, consistency did not differ between high and low confidence for trials followed by informative feedback: Neither confidence and accuracy interaction (LME: Estimate=-0.01, CI=[-0.13 -0.11], χ^2 =0.052, p=0.5) nor confidence effects (LME: Estimate=0.01, CI=[-0.09 –0.11], χ^2 =0.03, p=0.8) were significant (Figure 20 and Section B.3 Table 6).

B.1.4 *Controlling for late changes of decision.*

We addressed the potential misclassification by our BCI of participant's decision on certain trials, which would have led to the wrong feedback being displayed. Indeed, BCI classification was sometimes inaccurate as underlined by the cross-fold validation procedure following the model construction phase (average model accuracy over 3 seconds window = 80.1%, SD= 12.3%). To improve BCI accuracy, the preferentially attended item for each trial was the item showing the highest correlation scores averaged over the 6 last output correlations before the feedback presentation. This allows eventual classification error to be compensated by a majority of correct output over this period. To ensure the robustness of individual correlation scores, each of them was based on correlation between reconstructed signals and the face and spiral target signals over a 3 seconds sliding window (see Figure 6 and Methods in Chapter 3). However, with this method, participant's change of decision lag to be correctly decoded since the 3 s window might contain a majority of signals related to the previous decision.

We identified late change of decision as change of decision occurring in the last 1,5 second before feedback presentation as it is the time-window we used to determine the current target of the participant. To account for such late change of decision, we looked at the evolution of the difference between the correlation score of the item reported as the decision minus the correlation score of the other item. We thus look at the slope of these signed differences over the last 1,5 second of the trial. We use the most liberal criterion possible as each trial with negative slope was considered to reflect a change of mind.

We fit a logistic model to the number of change of decision with Confidence, Accuracy and Feedback nature as fixed effect and the intercept per participants as random effect. Results are detailed in Table 7. As the distribution of feedback nature was orthogonal to de-

coding performance, the number of change of decision was evenly distributed between trial where informative (M = 0.060; SD= 0,02) and deceptive (M=0.069; SD=0,02) feedback were presented (GLME: OR=0.95, CI=[0.7-1.28],χ^2 =0.8, p=0.4). As we suspected erroneous trials contain a larger share of late decision change (GLME: OR=0,56, CI=[0,4-0,7], χ^2 =2.2, p<0.0001) (post-hoc paired t-test: t(58)=-5.09, p<0,0001). Moreover, we found more decision change in trials reported with high confidence (M=0,11, SD=0,06) than trials reported with low confidence (M=0,07, SD=0,07) (paired t-test: t(58)=3,00, p=0.004). Thus, we remove from our analysis the data potentially containing undetected change of mind.

B.1.4.1 *Exogenous influence on confidence.*

Since deceptive trials reported with high confidence present lower accuracy than those reported with low confidence, we suggest that confidence is impacted by external cues. An alternative hypothesis is that high confidence has been attributed to trials with undetected late change of decision, inaccurately classified as confabulation. We did not find a significantly higher rate of undetected late change when participants report high confidence for their confabulation (paired t test t=-1,54, p=0,13). Although this difference failed to reach a significant threshold, we nonetheless repeat our previous analysis excluding trials containing a late change of decision.

Deceptive feedback still overturns the classical relationship between confidence and accuracy (GLME: OR= 3.46, CI= [2.29-4.01], χ^2 = 235.8, p<0.0001, See supplementary Table 8). As shown previously, we found a positive correlation by which confidence increased when accuracy increased: (high confidence: M = 0.81, SE = 0.03; low confidence: M = 0.58, SE = 0.03), z=-5,3, d=-1,49, p<0.0001 signed-rank test. Moreover, when participants received a deceptive feedback, this correlation was inverted by confidence rising up as accuracy decreased: (high confidence: M=0.48, SD=0.04; low confidence: M = 0.66, SD = 0.03), z=-4,5, d=0,93, p<0.0001 signed-rank test. These results reinforced our original assumption that external cues not only influence the content of introspection but also the confidence participants have in this introspective process.

B.1.4.2 *Internal decision evidence consistency modulate the impact of external cues on confidence.*

If a change of decision occurs during the 1.5 second preceding the feedback, some correlation scores will favour alternatively one and the other item. It results that the internal decision evidence computed as the accumulated difference between the correlation scores will be artificially lowered. Consistency, since proportional to internal decision evidence will be impacted in the same way. Therefore, we need to confirm that confabulations reported with high confidence were actually supported by less consistent decision evidence and not the

mere result of undetected change of mind.

To do this, we model consistency using accuracy and confidence after excluding trials containing late change of decision (see Table 9 in Section B.3). Relationship between internal decision evidence consistency and confidence report was still different for accurate and confabulated introspection (LME: Estimate=0.19, CI=[0.07 0.31], χ^2 =6.9 , p<0.01). Indeed, consistency was inversely correlated with confidence for confabulated reports (Low confidence: M = 0.19, SEM= 0.07; High confidence M=-0.03, SEM= 0.06), signed rank test z=-1,98, d=0.64, p<0,05. Again, consistency of internal evidence supporting accurate introspections fail to present a clear increase with confidence as we suggest that a ceiling effect prevent consistency to grow further (low confidence: M=0.38 SEM=0.05; high confidence: M=0.47, SEM= 0.07; paired t-test t=-1.1, d=0,24, p=0,27). These results confirm our original interpretation and show that absence of detection by our BCI of late change of mind does not account for the weaker internal evidence consistency associated with confident confabulation.

B.2 FIGURES

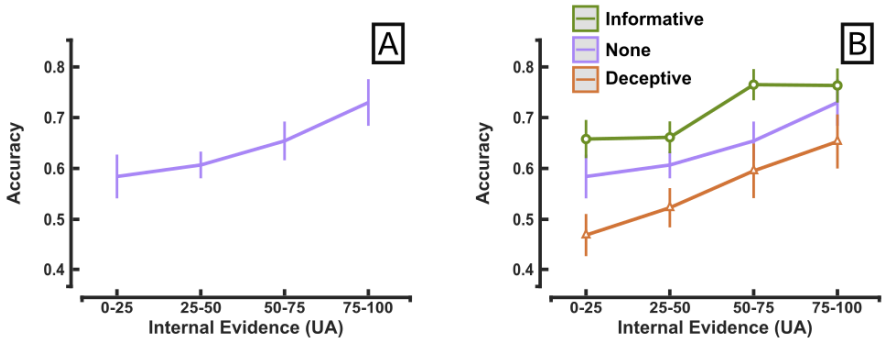

Figure 19: A.Decisions followed by no outcome: Accuracy in y-axis as percentage of correct trial . For each participant, we look at the distribution of internal evidence across trials. We group trials having their internal evidence between the 5th and 25th percentiles (1st point), the 25th and 50th percentiles(2nd point), the 50th and 75th percentiles (3rd point) and the 75thand 95th percentiles (4th point) respectively. Within each group for each participant we average the accuracy. Vertical bars represent across participants 1 000 times bootstrapped confidence intervals. B. Same plot for decision followed by either a deceptive (orange), informative (green) or no outcome (violet).

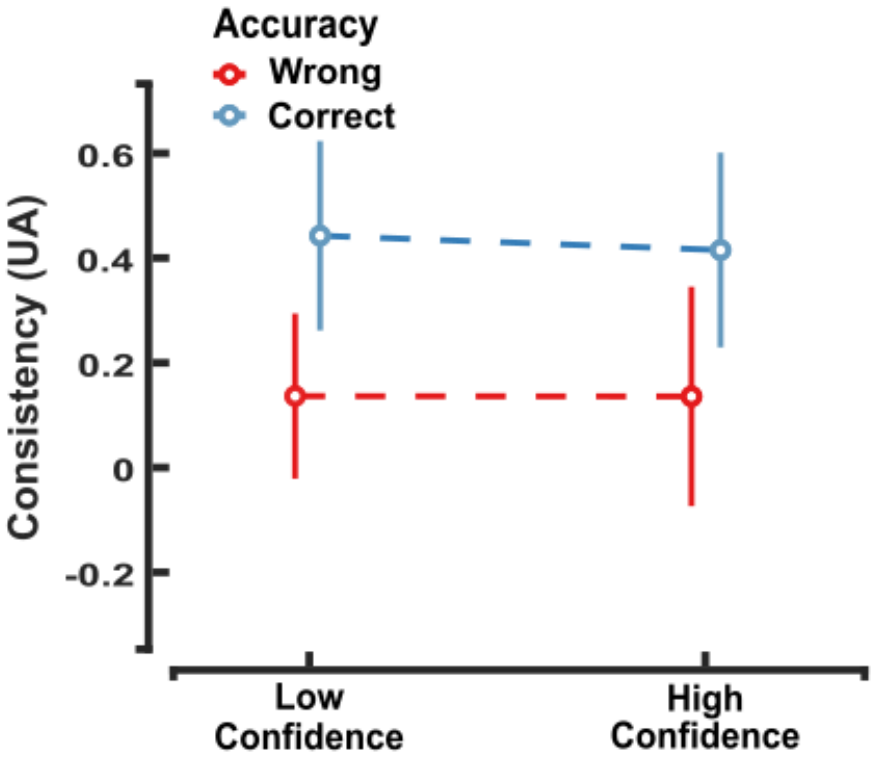

Figure 20: Consistency of internal decision evidence is shown for correct (blue) and incorrect (red) introspective report following the presentation of informative feedback. Reports are separated given the confidence attributed by participants. Vertical bars represent across-participants standard error to the mean.

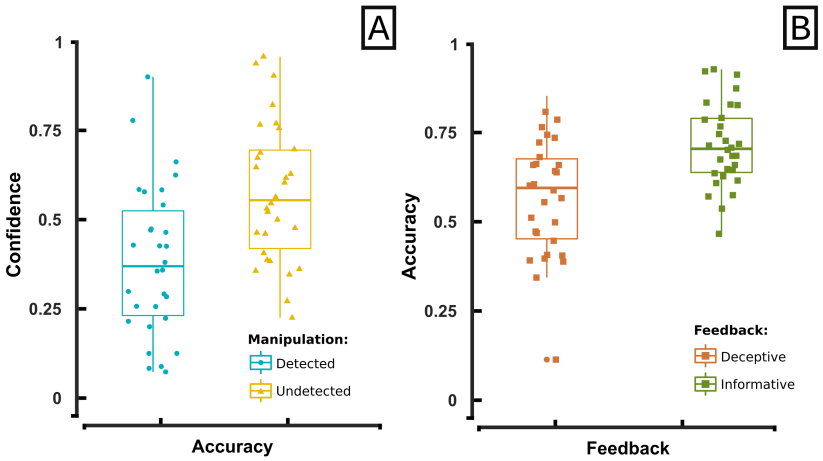

Figure 21: (A) Confidence as a function of manipulation detection: Confidence was coded as a binary variable: low confidence was labelled as 0 and High confidence was labelled as 1. We average confidence acros trials ended by a deceptive feedback respectively for cases where this feedback manipulation went detected (blue) or undetected (yellow). (B) Introspective reports accuracy regarding the feedback: The plot shows the accuracy in introspective reports in the y-axis averaged by participants. Trials followed by informative feedback are represented in green. Trials followed by deceptive feedback are represented in orange and indicated how much participants can detect feedback manipulation.

B.3 METHODS AND TABLES

To assess the respective influences of internal decision evidence and external cues on introspective reports, accuracy was modelled using generalized mixed-effect model with IE, Feedback and their interaction as fixed effect and Subject as random effect. Model was fitted to data of the experimental phase, with Feedback coded as deceptive or informative.

The model used was:

$$Accuracy \sim IE \times Feedback + (1|Subject). \tag{6}$$

The same model was fitted to data including both experimental and control phases. This time, Feedback was coded as deceptive, informative or none. The model used was:

$$Accuracy \sim IE \times Feedback + (1|Subject). \tag{7}$$

To measures the effect of external cues on reported confidence, Accuracy was modelled using generalized mixed-effect model with Confidence and Feedback as fixed effect and Subject as random effect following the original formula:

$$Accuracy \sim Confidence \times Feedback + (1|Subject). \tag{8}$$

Before we analysed the effect of internal decision evidence consistency on confabulation, we verified that accuracy of introspective reports do not separate confabulation from accurate reports the same way following informative and a deceptive feedback. We model log transformed consistency using a linear mixed effect model with Accuracy, Feedback and Confidence as fixed effects and Subject as random effect. The model used was:

$$Consistency \sim Accuracy \times Confidence \times Feedback + (1|Subject). \tag{9}$$

As we report a significant triple interaction, we then looked specifically at trials followed by deceptive feedback because in this condition incorrect reports correspond to confabulation. We do so by fitting linear mixed effect models to trials followed by deceptive feedback with log-transformed Consistency as dependant variable, Accuracy and Confidence as fixed effect and Subject as random effect. The model used was:

$$Consistency \sim Accuracy \times Confidence + (1|Subject). \tag{10}$$

We reproduced the former analyses for trials followed by informative feedback. The model used was:

$$Consistency \sim Accuracy \times Confidence + (1|Subject). \tag{11}$$

Accuracy

PREDICTORS	ODDS RATIOS	CI	Z-VALUE	P
(Intercept)	1.03	0.79 – 1.35	0.22	0.827
IE	1.19	1.13 – 1.26	6.34	<0.001
Feedback [Deceptive]	0.86	0.67 – 1.12	-1.12	0.264
Feedback [Informative]	1.55	1.19 – 2.02	3.27	0.001
IE * Feedback [Deceptive]	0.95	0.88 – 1.02	-1.47	0.143
IE * Feedback[Informative]	0.98	0.91 – 1.05	-0.66	0.508
RANDOM EFFECTS				
σ_2	3.29			
τ_{00} Subject	0.13			
ICC	0.04			
$N_{S}ubject$	16			
Observations	4911			
Marginal R^2 / Conditional R^2	0.059 / 0.096			

Table 1

As a sanity check, we modelled the occurrence of a late and potentially undecoded change of mind by fitting a generalized mixed-effect model to all our data using Confidence Feedback and Accuracy as fixed effect and Subject as a random effect. The model used was:

$$ChangeofMind \sim Confidence \times Feedback \times Accuracy + (1|Subject) \tag{12}$$

We then rerun analysis presented respectively in Table 3 and Table 5 to control that the observed effects were not attributable to a larger proportion of non decoded late change of mind in high confidence condition.

Accuracy

PREDICTORS	ODDS RATIOS	CI	Z-VALUE	P
(Intercept)	1.33	1.11 – 1.60	3.12	0.002
Confidence.L	0.61	0.55 – 0.68	-9.44	<0.001
Feedback [Informative]	1.93	1.74 – 2.15	12.34	<0.001
Confidence * Feedback	3.46	2.98 – 4.01	16.37	<0.001
RANDOM EFFECTS				
σ_2	3.29			
τ_{00} Subject	0.22			
ICC	0.06			
$N_{Subject}$	30			
Observations	6956			
Marginal R^2 / Conditional R^2	0.084 / 0.141			

Table 2

Consistency

PREDICTORS	ODDS RATIOS	CI	Z-VALUE	P
(Intercept)	0.18	0.09 – 0.27	3.95	<0.001
Accuracy [Correct]	0.33	0.25 – 0.42	7.64	<0.001
Feedback [Informative]	0.03	-0.07 – 0.13	0.58	0.564
Confidence.L	-0.11	-0.20 – -0.02	-2.37	0.018
Accuracy *Feedback	-0.02	-0.15 – 0.10	-0.38	0.705
Accuracy :Confidence	0.18	0.06 – 0.30	3.01	0.003
Feedback :Confidence	0.14	-0.00 – 0.29	1.96	0.050
Accuracy :Feedback :Confidence	-0.24	-0.42 – -0.06	-2.62	0.009
RANDOM EFFECTS				
σ_2	1.49			
τ_{00} Subject	0.03			
ICC	0.02			
$N_{Subject}$	30			
Observations	6956			
Marginal R^2 / Conditional R^2	0.017 / 0.038			

Table 3

Consistency

PREDICTORS	ODDS RATIOS	CI	Z-VALUE	P
(Intercept)	0.17	0.08 – 0.26	3.56	<0.001
Confidence.L	-0.11	-0.20 – -0.02	-2.29	0.022
Accuracy [Correct]	0.36	0.27 – 0.44	7.82	<0.001
Confidence * Accuracy	0.18	0.06 – 0.30	2.86	0.004
RANDOM EFFECTS				
σ_2	1.57			
τ_{00} Subject	0.03			
ICC	0.02			
N Subject	30			
Observations	3469			
Marginal R^2 / Conditional R^2	0.021 / 0.042			

Table 4

Consistency

PREDICTORS	ODDS RATIOS	CI	Z-VALUE	P
(Intercept)	0.21	0.10 – 0.31	3.81	<0.001
Confidence.L	0.03	-0.08 – 0.14	0.51	0.611
Accuracy [Correct]	0.32	0.22 – 0.41	6.73	<0.001
Confidence.L * Accuracy	-0.05	-0.18 – 0.08	-0.73	0.468
RANDOM EFFECTS				
σ_2	1.39			
τ_{00} Subject	0.04			
ICC	0.03			
$N_{Subject}$	30			
Observations	3487			
Marginal R^2 / Conditional R^2	0.014 / 0.042			

Table 5

Change of Mind

PREDICTORS	ODDS RATIOS	CI	Z-VALUE	P
(Intercept)	0.07	0.06 – 0.09	-24.64	<0.001
Confidence.L	1.37	1.02 – 1.84	2.07	0.038
Accuracy [Correct]	0.54	0.39 – 0.75	-3.76	<0.001
Feedback [Informative]	0.96	0.69 – 1.35	-0.21	0.835
Confidence * Accuracy	0.57	0.37 – 0.90	-2.42	0.016
Confidence * Feedback	0.63	0.39 – 1.02	-1.89	0.059
Accuracy * Feedback	0.87	0.54 – 1.41	-0.57	0.569
Confidence * Accuracy * Feedback	2.74	1.39 – 5.40	2.91	0.004
RANDOM EFFECTS				
σ_2	3.29			
τ_{00} Subject	0.00			
$N_{Subject}$	30			
Observations	6956			
Marginal R^2 / Conditional R^2	0.041 / NA			

Table 6

Accuracy

PREDICTORS	ODDS RATIOS	CI	Z-VALUE	P
(Intercept)	1.38	1.14 – 1.66	3.31	0.001
Feedback [Informative]	1.94	1.74 – 2.16	12.06	<0.001
Confidence.L	0.63	0.57 – 0.70	-8.62	<0.001
Feedback :Confidence	3.28	2.81 – 3.82	15.19	<0.001
RANDOM EFFECTS				
σ_2	3.29			
τ_{00}	0.24			
ICC	0.07			
N Subject	30			
Observations	6620			
Marginal R^2 / Conditional R^2	0.080 / 0.143			

Table 7

Consistency

PREDICTORS	ODDS RATIOS	CI	Z-VALUE	P
(Intercept)	0.23	0.13 – 0.33	4.64	<0.001
Confidence.L	-0.11	-0.21 – -0.02	-2.34	0.019
Accuracy [Correct]	0.33	0.24 – 0.42	7.14	<0.001
Confidence.L * Accuracy	0.18	0.05 – 0.30	2.75	0.006
RANDOM EFFECTS				
σ_2	1.55			
τ_{00} Subject	0.04			
ICC	0.02			
$N_{Subject}$	30			
Observations	3286			
Marginal R^2 / Conditional R^2	0.019 / 0.043			

Table 8

C

EXTENDED DATA FOR CHAPTER 4

C.1 RESULTS

C.1.1 *Number of random versus determined reports.*

Participants report an equal amount of determined (M=58.2, SEM=14.0) and random (M=65,6 SEM=12.4) choice in inclusive trials paired t-test t=-0.3, p=0.8, Cohen's d =0.17. Same was true for exclusive trials where determined choice (M=65.7,SEM=12.6) and random choice (M=56.6, SEM=8.9) were reported at comparable frequency (paired t-test t=-0.45, p=0.6, Cohen's d =0.27).

C.1.2 *AAS level does not reflect a successful decision phase.*

Noteworthy an alternative interpretation of those results could be that participants are mostly aware of the content of their deliberation but barely succeed to correctly attend the desired item in the post-cue period. AAS would therefore reflect a successful decision phase. To control for this alternative explanation, we look at the impact of AAS on participants' accuracy. We conduct this analysis both for experiment 1 and 2. In both case, AAS does affect the accuracy during exclusion trials (experiment 1: GLME, OR: 1.07, CI:[1-1.14],χ^2=4.2 , p<0.05 ; experiment 2: GLME, OR: 1.14, CI:[1.06-1.23],χ^2=11.5 , p<0.001; see Figure 22A). However, should AAS reflect the ability to focus on the desired item, we would expect it to also influence accuracy in inclusion trials. As shown in Figure 22B, that was neither the case in experiment 1 (GLME, OR: 0.95, CI:[0.89-1.01],χ^2=2.5 , p>0.1) nor in experiment 2 (GLME, OR: 1, CI:[0.92-1.08],$\chi^2 = 4.7 * 10^{-}5$, p=0.99) .

We complete our control analysis by looking at the impact of AAS on the relationship between accuracy and IB for inclusion trials reported as determined choices. Again, should AAS reflect the ability to focus on the desired item, we would expect it to also influence the effect of IB on accuracy in inclusive trials. We found no effect of the interaction between IB and AAS on accuracy (see Figure 22D) (GLMER, OR:1.11, CI:[0.72-1.71],χ^2=0.22 , p=0.6). Along with the good performance of our decoding model (M=76.2, SEM= 2.2) checked by cross-fold validation, we conclude that the effect of AAS on accuracy does not reflect a mere improvement in the use of the BCI setup.

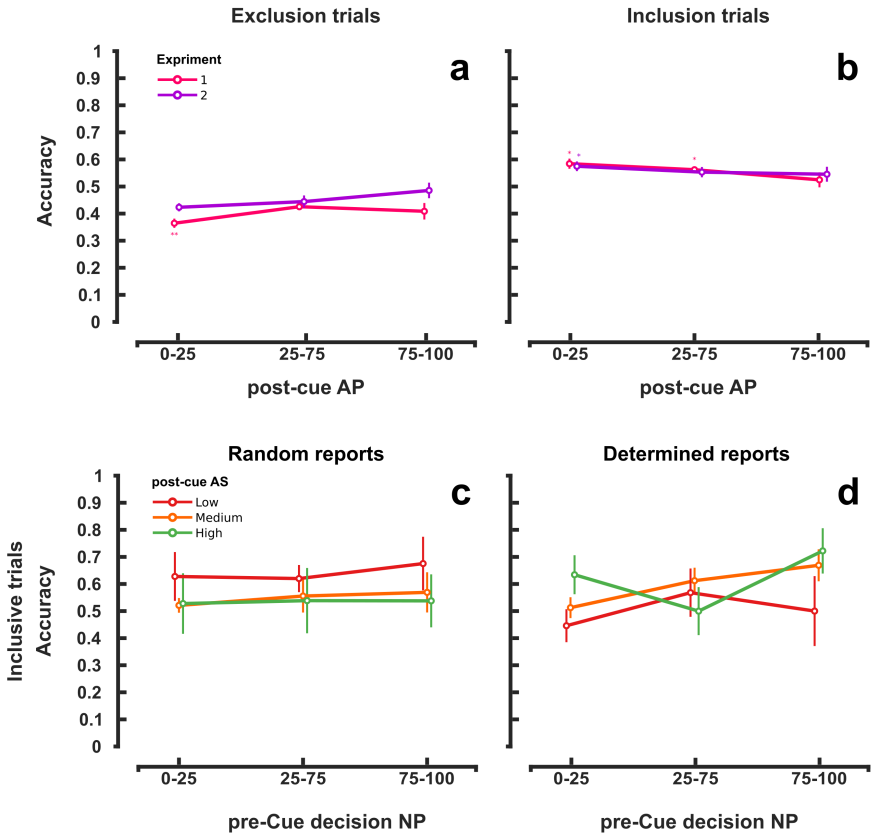

Figure 22: Accuracy does not reflect a mere success in using BCI setup. (a-b)Influence of post-cue attention allocation strength (AAS) on accuracy for experiment 1 (light purple) and 2 (dark purple) in exclusive (a) and exclusive (b) trials. For each participant, we computed the distribution of AAS across all trials in terms of percentile. (c-d)Impact of decision neural precursor (IB) on accuracy for different levels of post cue attention allocation strength (AAS) in inclusive trials. Trials were sorted following subjective report with random choice report (a) and determined choice report (b) respectively. Data are presented in the same way they are on Figure 12.

C.2 METHODS AND TABLES

To assess a potential effect of pre-cue deliberation on subsequent free decisions, we modelized accuracy using generalized mixed-effect model (GLME) with IB and the nature of the cue as fixed effect and the identity of participants as random effect (Table 9).

After reduction, the model used was:

$$\mathrm{Accuracy} \sim \mathrm{IB} \times \mathrm{Cue} + (1|\mathrm{Subject}). \tag{13}$$

Since the interaction terms reveal a different relationship between IB and accuracy in inclusive and exclusive trials, we run two separate GLME model for each type of cue (Table 10,Table 11).

Accuracy				
PREDICTORS	ODDS RATIOS	CI	Z-VALUE	P
(Intercept)	1.14	1.00 – 1.31	1.92	0.055
IB	1.26	1.08 – 1.47	2.95	0.003
Cue [Exclusive]	0.80	0.65 – 0.97	-2.24	0.025
IB * Cue[Exclusive]	0.67	0.54 – 0.84	-3.51	<0.001
RANDOM EFFECTS				
σ_2	3.29			
τ_{00} Subject	0.00			
$N_S ubject$	13			
Observations	3246			
Marginal R^2 / Conditional R^2	0.022 / NA			

Table 9: Experiment 1, all data.

After reduction, the model used was:

$$Accuracy \sim IB + (1|Subject). \tag{14}$$

The exact same analyses were repeated in our second experiment (Table 12,Table 13,Table 14):

We then investigate whether subjective reports reflected participants awareness by measuring the impact of reports on accuracy. We run the analyses for inclusive and exclusive trials separately since accuracy only in exclusive trial did accuracy reflects a conscious processing of early deliberations.

We first model the impact of IB on accuracy as a function of reported choice. We used a GLME with IB and report as fixed effect and participants as a random effect. (Inclusive trial Table 15. Exclusive trial Table 16)

After reduction, the model used was:

$$Accuracy \sim IB \times Report + (1|Subject). \tag{15}$$

To evaluate the difference of impact of IB on accuracy for report of random and determined choice, we separate our data depending of the participants' reports (Table 17,Table 18,Table 19,Table 20).

After reduction, the model used was:

$$Accuracy \sim IB + (1|Subject). \tag{16}$$

Accuracy

PREDICTORS	ODDS RATIOS	CI	Z-VALUE	P
(Intercept)	1.16	0.98 – 1.37	1.78	0.075
IB	1.27	1.08 – 1.48	2.95	0.003
RANDOM EFFECTS				
σ_2	3.29			
τ_{00} Subject	0.02			
ICC	0.01			
$N_S ubject$	13			
Observations	1622			
Marginal R^2 / Conditional R^2	0.007 / 0.014			

Table 10: Inclusve Cue

We then investigate the respective impact of early decision content and neural signal variability on participants' subjective reports. We used a Cumulative linked mixed effect model (CLMM) to regress the reports with IB and Variability as fixed effects and participants as random effect. The report were coded as follow: 1) Random choice high confidence, 2) Random choice low confidence, 3) Determined choice low confidence, 4) Determined choice high confidence (Table 21).

After reduction, the model used was:

$$Report \sim IB + Variability + (1|Subject). \tag{17}$$

Finally, we assess the effect of post-cue attentional allocation on awareness of early deliberation. We model using GLME the impact of AAS on the relationship between IB and accuracy. We conducted this analysis separately in inclusive and exclusive trials as only the latter track participants awareness (Table 22,Table 23).

After reduction, the model used was:

$$Accuracy \sim IB \times AAS + (1|Subject). \tag{18}$$

Accuracy				
PREDICTORS	ODDS RATIOS	CI	Z-VALUE	P
(Intercept)	0.91	0.77 – 1.08	-1.07	0.287
IB	0.84	0.71 – 0.99	-2.13	0.033
RANDOM EFFECTS				
σ_2	3.29			
τ_{00} Subject	0.02			
ICC	0.01			
$N_Subject$	13			
Observations	1624			
Marginal R^2 / Conditional R^2	0.003 / 0.010			

Table 11: Exclusive Cue.

Accuracy				
PREDICTORS	ODDS RATIOS	CI	Z-VALUE	P
(Intercept)	1.40	1.12 – 1.74	2.98	0.003
IB	1.28	1.14 – 1.45	4.05	<0.001
Cue [Exclusive]	0.75	0.64 – 0.88	-3.54	<0.001
IB * Cue[Exclusive]	0.69	0.59 – 0.82	-4.36	<0.001
RANDOM EFFECTS				
σ_2	3.29			
τ_{00} Subject	0.09			
ICC	0.03			
$N_Subject$	10			
Observations	2479			
Marginal R^2 / Conditional R^2	0.017 / 0.044			

Table 12: Experiment 2, all data

Accuracy

PREDICTORS	ODDS RATIOS	CI	Z-VALUE	P
(Intercept)	1.40	1.14 – 1.71	3.28	0.001
IB	1.28	1.14 – 1.45	4.04	<0.001
RANDOM EFFECTS				
σ_2	3.29			
τ_{00} Subject	0.07			
ICC	0.02			
$N_{S}ubject$	10			
Observations	1240			
Marginal R^2 / Conditional R^2	0.017 / 0.038			

Table 13: Inclusive Cue

Accuracy

PREDICTORS	ODDS RATIOS	CI	Z-VALUE	P
(Intercept)	1.04	0.84 – 1.29	0.36	0.722
IB	0.89	0.79 – 1.00	-2.05	0.041
RANDOM EFFECTS				
σ_2	3.29			
τ_{00} Subject	0.09			
ICC	0.03			
$N_{S}ubject$	10			
Observations	1239			
Marginal R^2 / Conditional R^2	0.004 / 0.030			

Table 14: Exclusive Cue

Accuracy				
PREDICTORS	ODDS RATIOS	CI	Z-VALUE	P
(Intercept)	1.42	1.17 – 1.72	3.55	<0.001
report.L	1.15	0.95 – 1.38	1.47	0.141
IB	1.31	1.16 – 1.48	4.24	<0.001
report.L * IB	1.16	0.97 – 1.38	1.62	0.105
RANDOM EFFECTS				
σ_2	3.29			
τ_{00} Subject	0.06			
ICC	0.02			
$N_{Subject}$	10			
Observations	1240			
Marginal R^2 / Conditional R^2	0.024 / 0.041			

Table 15: Inclusive Cue

Accuracy				
PREDICTORS	ODDS RATIOS	CI	Z-VALUE	P
(Intercept)	1.04	0.84 – 1.29	0.34	0.734
IB	0.88	0.79 – 0.99	-2.17	0.030
report.L	0.99	0.83 – 1.19	-0.09	0.931
IB : report.L	1.20	1.02 – 1.42	2.23	0.026
RANDOM EFFECTS				
σ_2	3.29			
τ_{00} Subject	0.08			
ICC	0.03			
$N_{Subject}$	10			
Observations	1239			
Marginal R^2 / Conditional R^2	0.009 / 0.034			

Table 16: Exclusive Cue

Accuracy				
PREDICTORS	ODDS RATIOS	CI	Z-VALUE	P
(Intercept)	1.25	1.07 – 1.45	2.79	0.005
IB	1.17	1.00 – 1.37	1.97	0.049
RANDOM EFFECTS				
σ_2	3.29			
τ_{00} Subject	0.00			
$N_{Subject}$	10			
Observations	656			
Marginal R^2 / Conditional R^2	0.008 / NA			

Table 17: Inclusive cue, Random choice.

Accuracy				
PREDICTORS	ODDS RATIOS	CI	Z-VALUE	P
(Intercept)	0.99	0.84 – 1.16	-0.14	0.885
IB	0.77	0.65 – 0.90	-3.13	0.002
RANDOM EFFECTS				
σ_2	3.29			
τ_{00} Subject	0.00			
$N_{Subject}$	10			
Observations	577			
Marginal R^2 / Conditional R^2	0.022 / NA			

Table 18: Exclusive cue, Random choice.

Accuracy				
PREDICTORS	ODDS RATIOS	CI	Z-VALUE	P
(Intercept)	1.54	1.14 – 2.07	2.82	0.005
IB	1.45	1.20 – 1.76	3.80	<0.001
RANDOM EFFECTS				
σ_2	3.29			
τ_{00} Subject	0.12			
ICC	0.04			
$N_S ubject$	9			
Observations	584			
Marginal R^2 / Conditional R^2	0.036 / 0.070			

Table 19: Inclusive cue, Determined choice.

Accuracy				
PREDICTORS	ODDS RATIOS	CI	Z-VALUE	P
(Intercept)	1.00	0.75 – 1.34	0.02	0.987
IB	1.00	0.86 – 1.17	0.03	0.972
RANDOM EFFECTS				
σ_2	3.29			
τ_{00} Subject	0.14			
ICC	0.04			
$N_S ubject$	10			
Observations	662			
Marginal R^2 / Conditional R^2	0.000 / 0.040			

Table 20: Exclusive cue, Determined choice.

Report

PREDICTORS	ODDS RATIOS	CI	Z-VALUE	P
1 | 2	0.40	0.20 – 0.79	-2.63	0.009
2 | 3	1.21	0.61 – 2.39	0.55	0.583
3 | 4	6.09	3.07 – 12.08	5.16	<0.001
IB	0.98	0.91 – 1.05	-0.54	0.590
Variability	0.87	0.78 – 0.96	-2.80	0.005
RANDOM EFFECTS				
σ_2	3.29			
τ_{00} Subject	1.21			
ICC	0.27			
$N_{Subject}$	10			
Observations	2479			
Marginal R^2 / Conditional R^2	0.003 / 0.271			

Table 21: All data

Accuracy				
PREDICTORS	ODDS RATIOS	CI	Z-VALUE	P
(Intercept)	0.89	0.60 – 1.30	-0.62	0.536
AAS	1.17	0.82 – 1.67	0.86	0.392
IB	0.67	0.45 – 0.98	-2.07	0.039
AAS * IB	1.70	1.16 – 2.49	2.72	0.007
RANDOM EFFECTS				
σ_2	3.29			
τ_{00} Subject	0.08			
ICC	0.02			
$N_{Subject}$	10			
Observations	667			
Marginal R^2 / Conditional R^2	0.054 / 0.075			

Table 22: Exclusive Cue

Accuracy				
PREDICTORS	ODDS RATIOS	CI	Z-VALUE	P
(Intercept)	0.93	0.62 – 1.39	-0.35	0.724
AAS	1.04	0.71 – 1.53	0.19	0.846
IB	1.73	1.11 – 2.69	2.43	0.015
AAS * IB	1.11	0.72 – 1.70	0.47	0.636
RANDOM EFFECTS				
σ_2	3.29			
τ_{00} Subject	0.06			
ICC	0.02			
$N_{Subject}$	9			
Observations	593			
Marginal R^2 / Conditional R^2	0.049 / 0.067			

Table 23: Inclusive Cue

BIBLIOGRAPHY

Adrian, Edgar D and Bryan HC Matthews (1934). "The interpretation of potential waves in the cortex." In: *The Journal of Physiology* 81.4, pp. 440–471.

Ales, Justin M, Faraz Farzin, Bruno Rossion, and Anthony M Norcia (2012). "An objective method for measuring face detection thresholds using the sweep steady-state visual evoked response." In: *Journal of vision* 12.10, pp. 18–18.

Andersen, SK and MM Müller (2010). "Behavioral performance follows the time course of neural facilitation and suppression during cued shifts of feature-selective attention." In: *Proceedings of the National Academy of Sciences* 107.31, pp. 13878–13882.

Arakawa, Kenji, Shozo Tobimatsu, Hiroyuki Tomoda, Jun-ichi Kira, and Motohiro Kato (1999). "The effect of spatial frequency on chromatic and achromatic steady-state visual evoked potentials." In: *Clinical Neurophysiology* 110.11, pp. 1959–1964.

Asai, Tomohisa (2017). "Know thy agency in predictive coding: Metamonitoring over forward modeling." In: *Consciousness and Cognition* 51, pp. 82–99.

Attwell, David and Simon B Laughlin (2001). "An energy budget for signaling in the grey matter of the brain." In: *Journal of Cerebral Blood Flow & Metabolism* 21.10, pp. 1133–1145.

Barlow, HB (1969). "Trigger features, adaptation and economy of impulses." In: *Information Processing in the Nervous System*. Springer, pp. 209–230.

Barlow, Horace B et al. (1961). "Possible principles underlying the transformation of sensory messages." In: *Sensory communication* 1, pp. 217–234.

Bates, Douglas, Deepayan Sarkar, Maintainer Douglas Bates, and L Matrix (2007). "The lme4 package." In: *R package version* 2.1, p. 74.

Blakemore, Sarah-Jayne, Daniel M Wolpert, and Christopher D Frith (2002). "Abnormalities in the awareness of action." In: *Trends in cognitive sciences* 6.6, pp. 237–242.

Block, Ned (1995). "On a confusion about a function of consciousness." In: *Behavioral and brain sciences* 18.2, pp. 227–247.

— (2007). "Consciousness, accessibility, and the mesh between psychology and neuroscience." In: *Behavioral and brain sciences* 30.5-6, p. 481.

— (2011). "Perceptual consciousness overflows cognitive access." In: *Trends in cognitive sciences* 15.12, pp. 567–575.

Blumenthal, Arthur L (2001). "A Wundt primer." In: *Wilhelm Wundt in history*. Springer, pp. 121–144.

Bode, Stefan, Anna Hanxi He, Chun Siong Soon, Robert Trampel, Robert Turner, and John-Dylan Haynes (2011). "Tracking the unconscious generation of free decisions using uitra-high field fMRI." In: *PloS one* 6.6, e21612.

Bogacz, Rafal, Eric Brown, Jeff Moehlis, Philip Holmes, and Jonathan D Cohen (2006). "The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks." In: *Psychological review* 113.4, p. 700.

Bona, Silvia and Juha Silvanto (2014). "Accuracy and confidence of visual short-term memory do not go hand-in-hand: behavioral and neural dissociations." In: *PLoS One* 9.3, e90808.

Boring, Edwin G (1953). "A history of introspection." In: *Psychological bulletin* 50.3, p. 169.

Bourgeois, Alexia, Leonardo Chelazzi, and Patrik Vuilleumier (2016). "How motivation and reward learning modulate selective attention." In: *Progress in brain research*. Vol. 229. Elsevier, pp. 325–342.

Brass, Marcel, Ariel Furstenberg, and Alfred R Mele (2019). "Why neuroscience does not disprove free will." In: *Neuroscience & Biobehavioral Reviews* 102, pp. 251–263.

Brass, Marcel and Patrick Haggard (2007). "To do or not to do: the neural signature of self-control." In: *Journal of Neuroscience* 27.34, pp. 9141–9145.

— (2008). "The what, when, whether model of intentional action." In: *The Neuroscientist* 14.4, pp. 319–325.

Brown, Harriet, Rick A Adams, Isabel Parees, Mark Edwards, and Karl Friston (2013). "Active inference, sensory attenuation and illusions." In: *Cognitive processing* 14.4, pp. 411–427.

Canna, Antonietta, Anna Prinster, Michele Fratello, Luca Puglia, Mario Magliulo, Elena Cantone, Maria Agnese Pirozzi, Francesco Di Salle, and Fabrizio Esposito (2019). "A low-cost open-architecture taste delivery system for gustatory FMRI and BCI experiments." In: *Journal of Neuroscience Methods* 311, pp. 1–12.

Carrasco, Marisa (2011). "Visual attention: The past 25 years." In: *Vision research* 51.13, pp. 1484–1525.

Carruthers, Peter (2009). "How we know our own minds: The relationship between mindreading and metacognition." In: *Behavioral and brain sciences* 32.2, p. 121.

— (2010). "Introspection: Divided and partly eliminated." In: *Philosophy and Phenomenological Research* 80.1, pp. 76–111.

— (2011). "Higher-order theories of consciousness." In: *The stanford encyclopedia of philosophy*.

Chalk, Matthew, Olivier Marre, and Gašper Tkačik (2018). "Toward a unified theory of efficient, predictive, and sparse coding." In: *Proceedings of the National Academy of Sciences* 115.1, pp. 186–191.

Chambon, Valerian and Patrick Haggard (2012). "Sense of control depends on fluency of action selection, not motor performance." In: *Cognition* 125.3, pp. 441–451.

Chang, Ying-Chao, Nai-Wen Guo, Chao-Ching Huang, Shan-Tair Wang, and Jing-Jane Tsai (2000). "Neurocognitive attention and behavior outcome of school-age children with a history of febrile convulsions: a population study." In: *Epilepsia* 41.4, pp. 412–420.

Chen, Xiaogang, Yijun Wang, Masaki Nakanishi, Xiaorong Gao, Tzyy-Ping Jung, and Shangkai Gao (2015). "High-speed spelling with

a noninvasive brain–computer interface." In: *Proceedings of the national academy of sciences* 112.44, E6058–E6067.

Christensen, Rune Haubo Bojesen (2015). "ordinal—regression models for ordinal data." In: *R package version* 28, p. 2015.

Clark, Andy (2015). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.

— (2017). "Predictions, precision, and agentive attention." In: *Consciousness and cognition* 56, pp. 115–119.

Clark, James J and Alan L Yuille (2013). *Data fusion for sensory information processing systems*. Vol. 105. Springer Science & Business Media.

Cohen, Marlene R and John HR Maunsell (2009). "Attention improves performance primarily by reducing interneuronal correlations." In: *Nature neuroscience* 12.12, p. 1594.

Conant, Roger C and W Ross Ashby (1970). "Every good regulator of a system must be a model of that system." In: *International journal of systems science* 1.2, pp. 89–97.

Costall, Alan (2006). "'Introspectionism'and the mythical origins of scientific psychology." In: *Consciousness and Cognition* 15.4, pp. 634–654.

Danziger, Kurt (1980). "The history of introspection reconsidered." In: *Journal of the History of the Behavioral Sciences* 16.3, pp. 241–262.

De Gardelle, Vincent, Jérôme Sackur, and Sid Kouider (2009). "Perceptual illusions in brief visual presentations." In: *Consciousness and cognition* 18.3, pp. 569–577.

Deco, Gustavo and Ranulfo Romo (2008). "The role of fluctuations in perception." In: *Trends in neurosciences* 31.11, pp. 591–598.

Dennett, Daniel C (1991). "Real patterns." In: *The journal of Philosophy* 88.1, pp. 27–51.

Desimone, Robert and John Duncan (1995). "Neural mechanisms of selective visual attention." In: *Annual review of neuroscience* 18.1, pp. 193–222.

Desmurget, Michel and Angela Sirigu (2009). "A parietal-premotor network for movement intention and motor awareness." In: *Trends in cognitive sciences* 13.10, pp. 411–419.

Dienes, Zoltán and Josef Perner (2007). "Executive control without conscious awareness: The cold control theory of hypnosis." In: *Hypnosis and conscious states: The cognitive neuroscience perspective*, pp. 293–314.

Engelbert, Mark and Peter Carruthers (2010). "Introspection." In: *Wiley Interdisciplinary Reviews: Cognitive Science* 1.2, pp. 245–253.

Ericson, K Anders and Herbert A Simon (1980). "Verbal reports as data." In: *Psychological review* 87.3, pp. 215–251.

Ernst, Marc O and Martin S Banks (2002). "Humans integrate visual and haptic information in a statistically optimal fashion." In: *Nature* 415.6870, pp. 429–433.

Evans, Nathan, Steven Gale, Aaron Schurger, and Olaf Blanke (2015). "Visual feedback dominates the sense of agency for brain-machine actions." In: *PloS one* 10.6, e0130019.

Farrell, Simon and Stephan Lewandowsky (2018). *Computational modeling of cognition and behavior*. Cambridge University Press.
Fiala, Brian, Shaun Nichols, and Peter Carruthers (2009). "Confabulation, confidence, and introspection." In: *Behavioral and Brain Sciences* 32.2, p. 144.
FitzGerald, Thomas HB, Rosalyn J Moran, Karl J Friston, and Raymond J Dolan (2015). "Precision and neuronal dynamics in the human posterior parietal cortex during evidence accumulation." In: *Neuroimage* 107, pp. 219–228.
Flavell, John H (1976). "Metacognitive aspects of problem solving." In: *The nature of intelligence*.
— (1979). "Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry." In: *American psychologist* 34.10, p. 906.
Fletcher, Paul C and Chris D Frith (2009). "Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia." In: *Nature Reviews Neuroscience* 10.1, pp. 48–58.
Fourneret, Pierre and Marc Jeannerod (1998). "Limited conscious monitoring of motor performance in normal subjects." In: *Neuropsychologia* 36.11, pp. 1133–1140.
Friston, Karl J, Rebecca Lawson, and Chris D Frith (2013). "On hyperpriors and hypopriors: comment on Pellicano and Burr." In: *Trends in cognitive sciences* 17.1, p. 1.
Friston, Karl (2008). "Hierarchical models in the brain." In: *PLoS Comput Biol* 4.11, e1000211.
— (2010). "The free-energy principle: a unified brain theory?" In: *Nature reviews neuroscience* 11.2, pp. 127–138.
— (2012). "Prediction, perception and agency." In: *International Journal of Psychophysiology* 83.2, pp. 248–252.
Friston, Karl, Philipp Schwartenbeck, Thomas FitzGerald, Michael Moutoussis, Tim Behrens, and Raymond J Dolan (2013). "The anatomy of choice: active inference and agency." In: *Frontiers in human neuroscience* 7, p. 598.
Friston, Karl, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, and Giovanni Pezzulo (2017). "Active inference: a process theory." In: *Neural computation* 29.1, pp. 1–49.
Frith, Chris D (2012). "The role of metacognition in human social interactions." In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 367.1599, pp. 2213–2223.
Frith, Christopher D, Sarah-Jayne Blakemore, and Daniel M Wolpert (2000). "Abnormalities in the awareness and control of action." In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 355.1404, pp. 1771–1788.
Furstenberg, Ariel, Assaf Breska, Haim Sompolinsky, and Leon Y Deouell (2015). "Evidence of change of intention in picking situations." In: *Journal of Cognitive Neuroscience* 27.11, pp. 2133–2146.
Garcia, Javier O, Ramesh Srinivasan, and John T Serences (2013). "Near-real-time feature-selective modulations in human cortex." In: *Current Biology* 23.6, pp. 515–522.

Gazzaniga, Michael S, Roger W Sperry, et al. (1967). "Language after section of the cerebral commissures." In: *Brain* 90.1, pp. 131–148.

Golub, Matthew D, Steven M Chase, Aaron P Batista, and M Yu Byron (2016). "Brain–computer interfaces for dissecting cognitive processes underlying sensorimotor control." In: *Current opinion in neurobiology* 37, pp. 53–58.

Gordon, Noam, Roger Koenig-Robert, Naotsugu Tsuchiya, Jeroen JA Van Boxtel, and Jakob Hohwy (2017). "Neural markers of predictive coding under perceptual uncertainty revealed with Hierarchical Frequency Tagging." In: *Elife* 6, e22749.

Gordon, Noam, Naotsugu Tsuchiya, Roger Koenig-Robert, and Jakob Hohwy (2019). "Expectation and attention increase the integration of top-down and bottom-up signals in perception through different pathways." In: *PLoS biology* 17.4, e3000233.

Gottlieb, Jacqueline and Puiu Balan (2010). "Attention as a decision in information space." In: *Trends in cognitive sciences* 14.6, pp. 240–248.

Graziano, Martin and Mariano Sigman (2009). "The spatial and temporal construction of confidence in the visual scene." In: *PLoS One* 4.3, e4909.

Grover, Sonja C (1982). "A re-evaluation of the introspection controversy: Additional considerations." In: *Journal of General Psychology* 106, p. 205.

Haggard, Patrick (2008). "Human volition: towards a neuroscience of will." In: *Nature Reviews Neuroscience* 9.12, pp. 934–946.

— (2017). "Sense of agency in the human brain." In: *Nature Reviews Neuroscience* 18.4, p. 196.

Haggard, Patrick, Sam Clark, and Jeri Kalogeras (2002). "Voluntary action and conscious awareness." In: *Nature neuroscience* 5.4, pp. 382–385.

Hall, Lars, Petter Johansson, and Thomas Strandberg (2012). "Lifting the veil of morality: Choice blindness and attitude reversals on a self-transforming survey." In: *PloS one* 7.9, e45457.

Hall, Lars, Petter Johansson, Betty Tärning, Sverker Sikström, and Thérèse Deutgen (2010). "Magic at the marketplace: Choice blindness for the taste of jam and the smell of tea." In: *Cognition* 117.1, pp. 54–61.

Hall, Lars, Thomas Strandberg, Philip Pärnamets, Andreas Lind, Betty Tärning, and Petter Johansson (2013). "How the polls can be both spot on and dead wrong: Using choice blindness to shift political attitudes and voter intentions." In: *PloS one* 8.4, e60554.

Harris, Kenneth D and Thomas D Mrsic-Flogel (2013). "Cortical connectivity and sensory coding." In: *Nature* 503.7474, pp. 51–58.

Hauser, Tobias U, Reto Iannaccone, Philipp Stämpfli, Renate Drechsler, Daniel Brandeis, Susanne Walitza, and Silvia Brem (2014). "The feedback-related negativity (FRN) revisited: new insights into the localization, meaning and network organization." In: *Neuroimage* 84, pp. 159–168.

Herrmann, Christoph S (2001). "Human EEG responses to 1–100 Hz flicker: resonance phenomena in visual cortex and their poten-

tial correlation to cognitive phenomena." In: *Experimental brain research* 137.3-4, pp. 346–353.
Heyes, Cecilia, Dan Bang, Nicholas Shea, Christopher D Frith, and Stephen M Fleming (2020). "Knowing ourselves together: The cultural origins of metacognition." In: *Trends in Cognitive Sciences*.
Hohwy, Jakob (2013). *The predictive mind*. Oxford University Press.
Huebner, David, Thibault Verhoeven, Klaus-Robert Mueller, Pieter-Jan Kindermans, and Michael Tangermann (2018). "Unsupervised learning for brain-computer interfaces based on event-related potentials: Review and online comparison [research frontier]." In: *IEEE Computational Intelligence Magazine* 13.2, pp. 66–77.
Hunt, Laurence T, WM Nishantha Malalasekera, Archy O de Berker, Bruno Miranda, Simon F Farmer, Timothy EJ Behrens, and Steven W Kennerley (2018). "Triple dissociation of attention and decision computations across prefrontal cortex." In: *Nature neuroscience* 21.10, pp. 1471–1481.
Huxley, Thomas Henry (1874). "On the hypothesis that animals are automata, and its history." In: *Collected essays* 1.
Imaizumi, Shu and Yoshihiko Tanno (2019). "Intentional binding coincides with explicit sense of agency." In: *Consciousness and Cognition* 67, pp. 1–15.
Jacoby, Larry L (1991). "A process dissociation framework: Separating automatic from intentional uses of memory." In: *Journal of memory and language* 30.5, pp. 513–541.
James, William (1890). *The principles of psychology New York*. Holt and company.
Johansson, Petter, Lars Hall, and Sverker Sikström (2008). "From change blindness to choice blindness." In: *Psychologia* 51.2, pp. 142–155.
Johansson, Petter, Lars Hall, Sverker Sikström, and Andreas Olsson (2005). "Failure to detect mismatches between intention and outcome in a simple decision task." In: *Science* 310.5745, pp. 116–119.
Johansson, Petter, Lars Hall, Sverker Sikström, Betty Tärning, and Andreas Lind (2006). "How something can be said about telling more than we can know: On choice blindness and introspection." In: *Consciousness and cognition* 15.4, pp. 673–692.
Kang, Yul HR, Frederike H Petzschner, Daniel M Wolpert, and Michael N Shadlen (2017). "Piercing of consciousness as a threshold-crossing operation." In: *Current Biology* 27.15, pp. 2285–2295.
Kant, Immanuel (1910). *Kants gesammelte Schriften: Abt. Werke (9v.)* Vol. 1. G. Reimer.
Kerous, Bojan, Filip Skola, and Fotis Liarokapis (2018). "EEG-based BCI and video games: a progress report." In: *Virtual Reality* 22.2, pp. 119–135.
Khalighinejad, Nima, Aaron Schurger, Andrea Desantis, Leor Zmigrod, and Patrick Haggard (2018). "Precursor processes of human self-initiated action." In: *Neuroimage* 165, pp. 35–47.
Khalighinejad, Nima, Elisa Brann, Alexander Dorgham, and Patrick Haggard (2019). "Dissociating cognitive and motoric precursors of human self-initiated action." In: *Journal of cognitive neuroscience* 31.5, pp. 754–767.

Kinchla, R A (1992). "Attention." In: *Annual Review of Psychology* 43.1, pp. 711–742.

Klobassa, Daniela S, Theresa M Vaughan, Peter Brunner, NE Schwartz, Jonathan R Wolpaw, Christa Neuper, and EW Sellers (2009). "Toward a high-throughput auditory P300-based brain–computer interface." In: *Clinical neurophysiology* 120.7, pp. 1252–1261.

Knill, David C and Alexandre Pouget (2004). "The Bayesian brain: the role of uncertainty in neural coding and computation." In: *TRENDS in Neurosciences* 27.12, pp. 712–719.

Koenig-Robert, Roger and Rufin VanRullen (2013). "SWIFT: a novel method to track the neural correlates of recognition." In: *Neuroimage* 81, pp. 273–282.

Kool, Wouter, Amitai Shenhav, and Matthew M Botvinick (2017). "Cognitive control as cost-benefit decision making." In:

Kouider, Sid and Emmanuel Dupoux (2004). "Partial awareness creates the "illusion" of subliminal semantic priming." In: *Psychological science* 15.2, pp. 75–81.

Kouider, Sid, Vincent De Gardelle, Jérôme Sackur, and Emmanuel Dupoux (2010). "How rich is consciousness? The partial awareness hypothesis." In: *Trends in cognitive sciences* 14.7, pp. 301–307.

Krajbich, Ian and Antonio Rangel (2011). "Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions." In: *Proceedings of the National Academy of Sciences* 108.33, pp. 13852–13857.

Ku, Yi-Chia, Chih-Hsin Muo, Chin-Shein Ku, Chao-Huei Chen, Wen-Yuan Lee, Ein-Yiao Shen, Yen-Jung Chang, and Chia-Hung Kao (2014). "Risk of subsequent attention deficit-hyperactivity disorder in children with febrile seizures." In: *Archives of disease in childhood* 99.4, pp. 322–326.

Kunar, Melina A, Derrick G Watson, Konstantinos Tsetsos, and Nick Chater (2017). "The influence of attention on value integration." In: *Attention, Perception, & Psychophysics* 79.6, pp. 1615–1627.

Kurtz, Phillipp, Katharine A Shapcott, Jochen Kaiser, Joscha T Schmiedt, and Michael C Schmid (2017). "The influence of endogenous and exogenous spatial attention on decision confidence." In: *Scientific reports* 7.1, pp. 1–9.

Labecki, Maciej, Rafal Kus, Alicja Brzozowska, Tadeusz Stacewicz, Basabdatta S Bhattacharya, and Piotr Suffczynski (2016). "Nonlinear origin of SSVEP spectra—a combined experimental and modeling study." In: *Frontiers in computational neuroscience* 10, p. 129.

Lau, Hakwan C and Richard E Passingham (2006). "Relative blindsight in normal observers and the neural correlate of visual consciousness." In: *Proceedings of the National Academy of Sciences* 103.49, pp. 18763–18768.

Lau, Hakwan C, Robert D Rogers, and Richard E Passingham (2007). "Manipulating the experienced onset of intention after action execution." In: *Journal of cognitive neuroscience* 19.1, pp. 81–90.

Legaspi, Roberto and Taro Toyoizumi (2019). "A Bayesian psychophysics model of sense of agency." In: *Nature communications* 10.1, pp. 1–11.

Leibniz, Gottfried Wilhelm (1765). *Nouveaux essais sur l'entendement humain par l'auteur du système de l'harmonie préetablie*. Schreuder.

Lennie, Peter (2003). "The cost of cortical computation." In: *Current biology* 13.6, pp. 493–497.

Lenth, Russell, Henrik Singmann, Jonathon Love, Paul Buerkner, and Maxime Herve (2018). "Emmeans: Estimated marginal means, aka least-squares means." In: *R package version* 1.1, p. 3.

Li, Yuanqing, Huiqi Li, Cuntai Guan, and Zhengyang Chin (2007). "A self-training semi-supervised support vector machine algorithm and its applications in brain computer interface." In: *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*. Vol. 1. IEEE, pp. I–385.

Libet, Benjamin, Curtis A Gleason, Elwood W Wright, and Dennis K Pearl (1983). "Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential)." In: *Brain*. 106. Oxford University Press, pp. 623–642.

Locke, John (1690). *An essay concerning human understanding*. Kay & Troutman, 1847.

MacIntyre, Tadhg E, Eric R Igou, Mark J Campbell, Aidan P Moran, and James Matthews (2014). "Metacognition and action: a new pathway to understanding social and cognitive aspects of expertise in sport." In: *Frontiers in Psychology* 5, p. 1155.

Maoz, Uri, Ueli Rutishauser, Soyoun Kim, Xinying Cai, Daeyeol Lee, and Christof Koch (2013). "Predeliberation activity in prefrontal cortex and striatum and the prediction of subsequent value judgment." In: *Frontiers in neuroscience* 7, p. 225.

McFarland, DJ and JR Wolpaw (2017). "EEG-based brain–computer interfaces." In: *current opinion in Biomedical Engineering* 4, pp. 194–200.

McFarland, Dennis J, Janis Daly, Chadwick Boulay, and Muhammad A Parvaz (2017). "Therapeutic applications of BCI technologies." In: *Brain-Computer Interfaces* 4.1-2, pp. 37–52.

McGhie, Andrew and James Chapman (1961). "Disorders of attention and perception in early schizophrenia." In: *British Journal of Medical Psychology* 34.2, pp. 103–116.

McLaughlin, Owen and Jason Somerville (2013). "Choice blindness in financial decision making." In: *Judgment and Decision Making* 8.5, p. 577.

Mele, Alfred R (2014). *Free: Why science hasn't disproved free will*. Oxford University Press.

Mele, Alfred R, H William, et al. (1992). *Springs of action: Understanding intentional behavior*. Oxford University Press on Demand.

— (2009). *Effective intentions: The power of conscious will*. Oxford University Press on Demand.

Meyniel, Florent, Mariano Sigman, and Zachary F Mainen (2015). "Confidence as Bayesian probability: From neural origins to behavior." In: *Neuron* 88.1, pp. 78–92.

Miller, Daniel R (1962). "The study of social relationships: situation, identity, and social interaction." In:

Mitchell, Jude F, Kristy A Sundberg, and John H Reynolds (2007). "Differential attention-dependent response modulation across cell classes in macaque visual area V4." In: *Neuron* 55.1, pp. 131–141.

— (2009). "Spatial attention decorrelates intrinsic activity fluctuations in macaque area V4." In: *Neuron* 63.6, pp. 879–888.

Moore, James W and Paul C Fletcher (2012). "Sense of agency in health and disease: a review of cue integration approaches." In: *Consciousness and cognition* 21.1, pp. 59–68.

Moore, James W, Daniel M Wegner, and Patrick Haggard (2009). "Modulating the sense of agency with external cues." In: *Consciousness and cognition* 18.4, pp. 1056–1064.

Moore, James and Patrick Haggard (2008). "Awareness of action: Inference and prediction." In: *Consciousness and cognition* 17.1, pp. 136–144.

Moore, Tirin and Marc Zirnsak (2017). "Neural mechanisms of selective visual attention." In: *Annual review of psychology* 68, pp. 47–72.

Morgan, ST, JC Hansen, and SA Hillyard (1996). "Selective attention to stimulus location modulates the steady-state visual evoked potential." In: *Proceedings of the National Academy of Sciences* 93.10, pp. 4770–4774.

Moshman, David (2014). *Epistemic cognition and development: The psychology of justification and truth*. Psychology Press.

Mulder, MJ, L Van Maanen, and BU Forstmann (2014). "Perceptual decision neurosciences–a model-based review." In: *Neuroscience* 277, pp. 872–884.

Müller, MM, S Andersen, NJ Trujillo, P Valdes-Sosa, P Malinowski, and SA Hillyard (2006). "Feature-selective attention enhances color signals in early visual areas of the human brain." In: *Proceedings of the National Academy of Sciences* 103.38, pp. 14250–14254.

Mumford, David (1992). "On the computational architecture of the neocortex." In: *Biological cybernetics* 66.3, pp. 241–251.

Murakami, Masayoshi, M Inês Vicente, Gil M Costa, and Zachary F Mainen (2014). "Neural antecedents of self-initiated actions in secondary motor cortex." In: *Nature neuroscience* 17.11, pp. 1574–1582.

Neisser, Ulric and Robert Becklen (1975). "Selective looking: Attending to visually specified events." In: *Cognitive psychology* 7.4, pp. 480–494.

Nelson, Thomas O and Louis Narens (1990). "Metamemory: A theoretical framework and new findings." In: *The psychology of learning and motivation* 26, pp. 125–141.

Nisbett, Richard E and Stanley Schachter (1966). "Cognitive manipulation of pain." In: *Journal of Experimental Social Psychology* 2.3, pp. 227–236.

Nisbett, Richard E and Timothy D Wilson (1977). "Telling more than we can know: verbal reports on mental processes." In: *Psychological review* 84.3, p. 231.

Norcia, Anthony M, L Gregory Appelbaum, Justin M Ales, Benoit R Cottereau, and Bruno Rossion (2015). "The steady-state visual

evoked potential in vision research: a review." In: *Journal of vision* 15.6, pp. 4–4.

O'Regan, J Kevin and Alva Noë (2001). "A sensorimotor account of vision and visual consciousness." In: *Behavioral and brain sciences* 24.5, p. 939.

O'Sullivan, James A, Alan J Power, Nima Mesgarani, Siddharth Rajaram, John J Foxe, Barbara G Shinn-Cunningham, Malcolm Slaney, Shihab A Shamma, and Edmund C Lalor (2015). "Attentional selection in a cocktail party environment can be decoded from single-trial EEG." In: *Cerebral cortex* 25.7, pp. 1697–1706.

Oliveira, Flavio TP, John J McDonald, and David Goodman (2007). "Performance monitoring in the anterior cingulate is not all error related: expectancy deviation and the representation of action-outcome associations." In: *Journal of cognitive neuroscience* 19.12, pp. 1994–2004.

Oostenveld, Robert, Pascal Fries, Eric Maris, and Jan-Mathijs Schoffelen (2011). "FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data." In: *Computational intelligence and neuroscience* 2011.

Orquin, Jacob L and Simone Mueller Loose (2013). "Attention and choice: A review on eye movements in decision making." In: *Acta psychologica* 144.1, pp. 190–206.

Overgaard, Morten (2015). "The challenge of measuring consciousness." In: *Behavioural methods in consciousness research*, pp. 7–19.

Overgaard, Morten, Peter Fazekas, et al. (2016). "Can no-report paradigms extract true correlates of consciousness." In: *Trends Cogn. Sci* 20, pp. 241–242.

Pacherie, Elisabeth (2008). "The phenomenology of action: A conceptual framework." In: *Cognition* 107.1, pp. 179–217.

Pares Pujolras, Elisabeth (2019). "Endogenicity and awareness in voluntary action." PhD thesis. UCL (University College London).

Parés-Pujolràs, Elisabeth, Yong-Wook Kim, Chang-Hwan Im, and Patrick Haggard (2019). "Latent awareness: Early conscious access to motor preparation processes is linked to the readiness potential." In: *Neuroimage* 202, p. 116140.

Passingham, RE (1987). "Two cortical systems for directing movement." In: *Ciba Found Symp*. Vol. 132. Wiley Online Library, pp. 151–164.

Passingham, Richard E, Sara L Bengtsson, and Hakwan C Lau (2010). "Medial frontal cortex: from self-generated action to reflection on one's own performance." In: *Trends in cognitive sciences* 14.1, pp. 16–21.

Pei, Francesca, Mark W Pettet, and Anthony M Norcia (2002). "Neural correlates of object-based attention." In: *Journal of Vision* 2.9, pp. 1–1.

Perez-Orive, Javier, Ofer Mazor, Glenn C Turner, Stijn Cassenaer, Rachel I Wilson, and Gilles Laurent (2002). "Oscillations and sparsening of odor representations in the mushroom body." In: *Science* 297.5580, pp. 359–365.

Pestilli, Franco, Marisa Carrasco, David J Heeger, and Justin L Gardner (2011). "Attentional enhancement via selection and pooling of early sensory responses in human visual cortex." In: *Neuron* 72.5, pp. 832–846.

Petersen, Steven E and Michael I Posner (2012). "The attention system of the human brain: 20 years after." In: *Annual review of neuroscience* 35, pp. 73–89.

Petitmengin, Claire, Anne Remillieux, Béatrice Cahour, and Shirley Carter-Thomas (2013). "A gap in Nisbett and Wilson's findings? A first-person access to our cognitive processes." In: *Consciousness and cognition* 22.2, pp. 654–669.

Picard, Fabienne and Karl Friston (2014). "Predictions, perception, and a sense of self." In: *Neurology* 83.12, pp. 1112–1118.

Posner, Michael I and Steven E Petersen (1990). "The attention system of the human brain." In: *Annual review of neuroscience* 13.1, pp. 25–42.

Posner, Michael I, Steven E Petersen, Peter T Fox, and Marcus E Raichle (1988). "Localization of cognitive operations in the human brain." In: *Science* 240.4859, pp. 1627–1631.

Pouget, Alexandre, Jan Drugowitsch, and Adam Kepecs (2016). "Confidence and certainty: distinct probabilistic quantities for different goals." In: *Nature neuroscience* 19.3, p. 366.

Qiu, Lirong, Jie Su, Yinmei Ni, Yang Bai, Xuesong Zhang, Xiaoli Li, and Xiaohong Wan (2018). "The neural system of metacognition accompanying decision-making in the prefrontal cortex." In: *PLoS biology* 16.4, e2004037.

Quiroga, R Quian, Leila Reddy, Gabriel Kreiman, Christof Koch, and Itzhak Fried (2005). "Invariant visual representation by single neurons in the human brain." In: *Nature* 435.7045, pp. 1102–1107.

Rahnev, Dobromir, Brian Maniscalco, Tashina Graves, Elliott Huang, Floris P De Lange, and Hakwan Lau (2011). "Attention induces conservative subjective biases in visual perception." In: *Nature neuroscience* 14.12, pp. 1513–1515.

Ransom, Madeleine, Sina Fazelpour, and Christopher Mole (2017). "Attention in the predictive mind." In: *Consciousness and cognition* 47, pp. 99–112.

Rao, Rajesh PN and Dana H Ballard (1999). "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects." In: *Nature neuroscience* 2.1, pp. 79–87.

Ratcliff, Roger and Gail McKoon (2008). "The diffusion decision model: theory and data for two-choice decision tasks." In: *Neural computation* 20.4, pp. 873–922.

Ratcliff, Roger, Philip L Smith, Scott D Brown, and Gail McKoon (2016). "Diffusion decision model: Current issues and history." In: *Trends in cognitive sciences* 20.4, pp. 260–281.

Raymond, Jennifer L and Javier F Medina (2018). "Computational principles of supervised learning in the cerebellum." In: *Annual review of neuroscience* 41, pp. 233–253.

Regan, D (1973). "Rapid objective refraction using evoked brain potentials." In: *Investigative Ophthalmology & Visual Science* 12.9, pp. 669–679.

Regan, David (1966). "Some characteristics of average steady-state and transient responses evoked by modulated light." In: *Electroencephalography and clinical neurophysiology* 20.3, pp. 238–248.

Reyes, Gabriel (2015). "Introspection of complex cognitive processes." PhD thesis.

Reyes, Gabriel, Carlos Schmidt, Nicolas Marchant, Vincent de Gardelle, Jaime R. Silva, and Jérôme Sackur (2018). "Blind Trust or Choice Blindness? Cues for Experimenter Competence Influence Choice Blindness." In: Publisher: OSF. URL: https://osf.io/ht769/ (visited on 05/20/2020).

Rezeika, Aya, Mihaly Benda, Piotr Stawicki, Felix Gembler, Abdul Saboor, and Ivan Volosyak (2018). "Brain–computer interface spellers: A review." In: *Brain sciences* 8.4, p. 57.

Ridderinkhof, K Richard, Markus Ullsperger, Eveline A Crone, and Sander Nieuwenhuis (2004). "The role of the medial frontal cortex in cognitive control." In: *science* 306.5695, pp. 443–447.

Rieznik, Andrés, Lorena Moscovich, Alan Frieiro, Julieta Figini, Rodrigo Catalano, Juan Manuel Garrido, Facundo Álvarez Heduan, Mariano Sigman, and Pablo A Gonzalez (2017). "A massive experiment on choice blindness in political decisions: Confidence, confabulation, and unconscious detection of self-deception." In: *PloS one* 12.2, e0171108.

Robinson, William (2019). "Epiphenomenalism." In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2019. Metaphysics Research Lab, Stanford University.

Roy, Jefferson E, Timothy J Buschman, and Earl K Miller (2014). "PFC neurons reflect categorical decisions about ambiguous stimuli." In: *Journal of cognitive neuroscience* 26.6, pp. 1283–1291.

Sackur, Jérôme (2009). "L'introspection en psychologie expérimentale." In: *Revue d'histoire des sciences* 62.2, pp. 349–372.

Saenz, Melissa, Giedrius T Buracas, and Geoffrey M Boynton (2002). "Global effects of feature-based attention in human visual cortex." In: *Nature neuroscience* 5.7, pp. 631–632.

Sato, Atsushi and Asako Yasuda (2005). "Illusion of sense of self-agency: discrepancy between the predicted and actual sensory consequences of actions modulates the sense of self-agency, but not the sense of self-ownership." In: *Cognition* 94.3, pp. 241–255.

Sauerland, Melanie, Anna Sagana, Henry Otgaar, and Nick J Broers (2014). "Self-relevance does not moderate choice blindness in adolescents and children." In: *PloS one* 9.6, e98563.

Schultze-Kraft, Matthias, Daniel Birman, Marco Rusconi, Carsten Allefeld, Kai Görgen, Sven Dähne, Benjamin Blankertz, and John-Dylan Haynes (2016). "The point of no return in vetoing self-initiated movements." In: *Proceedings of the National Academy of Sciences* 113.4, pp. 1080–1085.

Schultze-Kraft, Matthias, Elisabeth Parés-Pujolràs, Karla Matić, Patrick Haggard, and John-Dylan Haynes (2020). "Preparation and exe-

cution of voluntary action both contribute to awareness of intention." In: *Proceedings of the Royal Society B* 287.1923, p. 20192928.

Schurger, Aaron (2018). "Specific relationship between the shape of the readiness potential, subjective decision time, and waiting time predicted by an accumulator model with temporally autocorrelated input noise." In: *Eneuro* 5.1.

Schurger, Aaron, Jacobo D Sitt, and Stanislas Dehaene (2012). "An accumulator model for spontaneous neural activity prior to self-initiated movement." In: *Proceedings of the National Academy of Sciences* 109.42, E2904–E2913.

Schurger, Aaron, Steven Gale, Olivia Gozel, and Olaf Blanke (2017). "Performance monitoring for brain-computer-interface actions." In: *Brain and cognition* 111, pp. 44–50.

Searle, John R (1990a). "Cognitive science and the computer metaphor." In: *Artifical Intelligence, Culture and Language: On Education and Work*. Springer, pp. 23–34.

— (1990b). "Consciousness, explanatory inversion, and cognitive science." In: *Behavioral and Brain Sciences* 13.4, pp. 585–596.

Serences, John T (2011). "Mechanisms of selective attention: response enhancement, noise reduction, and efficient pooling of sensory responses." In: *Neuron* 72.5, pp. 685–687.

Sergent, Claire and Stanislas Dehaene (2004). "Is consciousness a gradual phenomenon? Evidence for an all-or-none bifurcation during the attentional blink." In: *Psychological science* 15.11, pp. 720–728.

Sergent, Claire, Valentin Wyart, Mariana Babo-Rebelo, Laurent Cohen, Lionel Naccache, and Catherine Tallon-Baudry (2013). "Cueing attention after the stimulus is gone can retrospectively trigger conscious perception." In: *Current biology* 23.2, pp. 150–155.

Shadlen, Michael N and William T Newsome (2001). "Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey." In: *Journal of neurophysiology* 86.4, pp. 1916–1936.

Shenhav, Amitai, Sebastian Musslick, Falk Lieder, Wouter Kool, Thomas L Griffiths, Jonathan D Cohen, and Matthew M Botvinick (2017). "Toward a rational and mechanistic account of mental effort." In: *Annual review of neuroscience* 40, pp. 99–124.

Shenoy, Pradeep, Matthias Krauledat, Benjamin Blankertz, Rajesh PN Rao, and Klaus-Robert Müller (2006). "Towards adaptive classification for BCI." In: *Journal of neural engineering* 3.1, R13.

Shepherd, Joshua (2015). "Conscious control over action." In: *Mind & Language* 30.3, pp. 320–344.

Somerville, Jason and Féidhlim McGowan (2016). "Can chocolate cure blindness? Investigating the effect of preference strength and incentives on the incidence of Choice Blindness." In: *Journal of Behavioral and Experimental Economics* 61, pp. 1–11.

Soon, Chun Siong, Marcel Brass, Hans-Jochen Heinze, and John-Dylan Haynes (2008). "Unconscious determinants of free decisions in the human brain." In: *Nature neuroscience* 11.5, pp. 543–545.

Soon, Chun Siong, Anna Hanxi He, Stefan Bode, and John-Dylan Haynes (2013). "Predicting free choices for abstract intentions."

In: *Proceedings of the National Academy of Sciences* 110.15, pp. 6217–6222.

Sperling, George (1960). "The information available in brief visual presentations." In: *Psychological monographs: General and applied* 74.11, p. 1.

Sprague, Thomas C, Sameer Saproo, and John T Serences (2015). "Visual attention mitigates information loss in small-and large-scale neural codes." In: *Trends in Cognitive Sciences* 19.4, pp. 215–226.

Srinivasan, Mandyam Veerambudi, Simon Barry Laughlin, and Andreas Dubs (1982). "Predictive coding: a fresh view of inhibition in the retina." In: *Proceedings of the Royal Society of London. Series B. Biological Sciences* 216.1205, pp. 427–459.

Strandberg, Thomas, David Sivén, Lars Hall, Petter Johansson, and Philip Pärnamets (2018). "False beliefs and confabulation can lead to lasting changes in political attitudes." In: *Journal of Experimental Psychology: General* 147.9, p. 1382.

Sully, James (1881). "Illusions of introspection." In: *Mind* 6.21, pp. 1–18.

Synofzik, Matthis, Gottfried Vosgerau, and Albert Newen (2008). "Beyond the comparator model: a multifactorial two-step account of agency." In: *Consciousness and cognition* 17.1, pp. 219–239.

Talmi, Deborah, Ryan Atkinson, and Wael El-Deredy (2013). "The feedback-related negativity signals salience prediction errors, not reward prediction errors." In: *Journal of Neuroscience* 33.19, pp. 8264–8269.

Talmi, Deborah, Lluis Fuentemilla, Vladimir Litvak, Emrah Duzel, and Raymond J Dolan (2012). "An MEG signature corresponding to an axiomatic model of reward prediction error." In: *Neuroimage* 59.1, pp. 635–645.

Team, R Core et al. (2012). "R: A language and environment for statistical computing. 2012." In: *Vienna, Austria: R Foundation for Statistical Computing* 10.

Torchiano, M (2016). *effsize: efficient effect size computation (R package)*.

Travers, Eoin, Nima Khalighinejad, Aaron Schurger, and Patrick Haggard (2020). "Do readiness potentials happen all the time?" In: *NeuroImage* 206, p. 116286.

Tsuchiya, Naotsugu, Melanie Wilke, Stefan Frässle, and Victor AF Lamme (2015). "No-report paradigms: extracting the true neural correlates of consciousness." In: *Trends in cognitive sciences* 19.12, pp. 757–770.

Ullmann-Margalit, Edna and Sidney Morgenbesser (1977). "Picking and choosing." In: *Social research*, pp. 757–785.

Ullsperger, Markus and D Yves Von Cramon (2004). "Neuroimaging of performance monitoring: error detection and beyond." In: *Cortex* 40.4-5, pp. 593–604.

Usher, Marius and James L McClelland (2001). "The time course of perceptual choice: the leaky, competing accumulator model." In: *Psychological review* 108.3, p. 550.

Vasser, Madis, Laurène Vuillaume, Axel Cleeremans, and Jaan Aru (2019). "Waving goodbye to contrast: self-generated hand move-

ments attenuate visual sensitivity." In: *Neuroscience of consciousness* 2019.1, niy013.

Vialatte, François-Benoît, Monique Maurice, Justin Dauwels, and Andrzej Cichocki (2010). "Steady-state visually evoked potentials: focus on essential paradigms and future perspectives." In: *Progress in neurobiology* 90.4, pp. 418–438.

Vidal, Jacques J (1973). "Toward direct brain-computer communication." In: *Annual review of Biophysics and Bioengineering* 2.1, pp. 157–180.

— (1977). "Real-time detection of brain events in EEG." In: *Proceedings of the IEEE* 65.5, pp. 633–641.

Vinje, William E and Jack L Gallant (2000). "Sparse coding and decorrelation in primary visual cortex during natural vision." In: *Science* 287.5456, pp. 1273–1276.

Vlassova, Alexandra, Chris Donkin, and Joel Pearson (2014). "Unconscious information changes decision accuracy but not confidence." In: *Proceedings of the National Academy of Sciences* 111.45, pp. 16214–16218.

Walter, Sven (2014). "Willusionism, epiphenomenalism, and the feeling of conscious will." In: *Synthese* 191.10, pp. 2215–2238.

Watson, John B (1913). "Psychology as the behaviorist views it." In: *Psychological review* 20.2, p. 158.

Wegner, Daniel M (2002). *The illusion of conscious will*. MIT press.

— (2003). "The mind's best trick: How we experience conscious will." In: *Trends in cognitive sciences* 7.2, pp. 65–69.

— (2004). "Precis of the illusion of conscious will (and commentaries and reply)." In:

Wegner, Daniel M, Betsy Sparrow, and Lea Winerman (2004). "Vicarious agency: experiencing control over the movements of others." In: *Journal of personality and social psychology* 86.6, p. 838.

Wegner, Daniel M and Thalia Wheatley (1999). "Apparent mental causation: Sources of the experience of will." In: *American psychologist* 54.7, p. 480.

Wheatley, Thalia and Jonathan Haidt (2005). "Hypnotic disgust makes moral judgments more severe." In: *Psychological science* 16.10, pp. 780–784.

Wierzchoń, Michał, Anna Anzulewicz, Justyna Hobot, Borysław Paulewicz, and Jérôme Sackur (2019). "In search of the optimal measure of awareness: Discrete or continuous?" In: *Consciousness and cognition* 75, p. 102798.

Wikipedia (2020). *Predictive coding — Wikipedia, The Free Encyclopedia*. http://en.wikipedia.org/w/index.php?title=Predictive%20coding&oldid=962401590. [Online; accessed 05-August-2020].

Wilimzig, Claudia, Naotsugu Tsuchiya, Manfred Fahle, Wolfgang Einhäuser, and Christof Koch (2008). "Spatial attention increases performance but not subjective confidence in a discrimination task." In: *Journal of vision* 8.5, pp. 7–7.

Wilson, Timothy D and Elizabeth W Dunn (2004). "Self-knowledge: Its limits, value, and potential for improvement." In: *Annual review of psychology* 55.

Wisniewski, David, Thomas Goschke, and John-Dylan Haynes (2016). "Similar coding of freely chosen and externally cued intentions in a fronto-parietal network." In: *Neuroimage* 134, pp. 450–458.

Wolpaw, Jonathan R, Dennis J McFarland, Gregory W Neat, and Catherine A Forneris (1991). "An EEG-based brain-computer interface for cursor control." In: *Electroencephalography and clinical neurophysiology* 78.3, pp. 252–259.

Wolpaw, Jonathan and Elizabeth Winter Wolpaw (2012). *Brain-computer interfaces: principles and practice*. OUP USA.

Wolpe, Noham, Patrick Haggard, Hartwig R Siebner, and James B Rowe (2013). "Cue integration and the perception of action in intentional binding." In: *Experimental brain research* 229.3, pp. 467–474.

Wundt, Wilhelm Max (1883). *Über psychologische Methoden*. W. Engelmann.

Wundt, Wilhelm (1888). "Selbstbeobachtung und innere Wahrnehmung." In: *Philosophische Studien* 4, pp. 292–309.

Yeung, Nick and Christopher Summerfield (2012). "Metacognition in human decision-making: confidence and error monitoring." In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 367.1594, pp. 1310–1321.

Yussen, Steven R and J Elizabeth Bird (1979). "The development of metacognitive awareness in memory, communication, and attention." In: *Journal of Experimental Child Psychology* 28.2, pp. 300–313.

Zhu, Danhua, Jordi Bieger, Gary Garcia Molina, and Ronald M Aarts (2010). "A survey of stimulation methods used in SSVEP-based BCIs." In: *Computational intelligence and neuroscience* 2010.

DECLARATION

I, Benjamin Rebouillat, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that the sources have been properly cited in the thesis.

Paris, September 2020

Benjamin Rebouillat

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

https://bitbucket.org/amiede/classicthesis/

Abstract

When we make a free choice, we feel conscious and in control of our decision processes. However, over the past decades, studies on introspection demonstrated that our self-knowledge faculties are crippled by illusory content. In Part i, we suggest that introspection can be framed as a hierarchically organized inference process and we proposed an innovative methodological approach to challenge this hypothesis. We used a free decision paradigm in which no high order nor low motor level processing were solicited. Further, we track in real time internal decision variables through a Brain Computer Interface (BCI), and probe both implicitly and explicitly participants' decision awareness. The present thesis investigates two main questions. First, what are the conditions for people to be aware of their impending decisions? Second, does people's introspections access genuine mental activity or are they pure retrospective illusions? Our results suggest that despite the general impression of a rich internal life, people are only partially aware of their impending decisions. If they can consciously track their upcoming decisions, they have no conscious access to those decisions' content. Yet, when recalling their recent choices, people can access internal representation of the chosen alternative. However, our results suggest that introspection has no privileged access to internal decision variables but rather stem from an integrative process involving both endogenous and exogenous cues. Introspective illusions thus reflect an imbalanced integration process, where weak and noisy internal variables are dominated by deceptive feedback. Overall, the present thesis provides new insights and methodological tools for the study of decision awareness emergence. Our results converge toward the idea that self-knowledge of decision is a hierarchically organized Bayesian inference process involving multiple cues.

Résumé

Nous concevons d'ordinaire nos choix comme conscients et sous notre contrôle. Toutefois, de nombreuses études montrent que nos processus introspectifs sont largement illusoires. Dans notre première partie, nous proposons que l'introspection peut être conceptualisée comme un processus d'inférence hiérarchique, et nous avançons une nouvelle approche pour en étudier les mécanismes sous-jacents. A cette fin, nous employons un protocole de prise de décision dans lequel les sujets ne peuvent accéder ni à leurs informations motrices, ni à des informations de haut niveau. En outre, nous mesurons les signaux neuronaux impliqués dans la prises de décision ainsi que la conscience que les sujets ont de leurs décisions. Cette thèse se penche sur deux questions: Premièrement sous quelles conditions peut-on être conscient de ses décisions? Deuxièmement pouvons nous accéder à nos processus mentaux par l'introspection, ou cette dernière n'est elle qu'une illusion? Nos résultats suggèrent, qu'en dépit d'un sentiment de richesse subjective, nous n'avons qu'un accès partiel aux contenus de nos décisions. Si l'on peut savoir qu'une décision est imminente, son contenu échappe à la conscience. Toutefois, les sujets peuvent accéder à une représentation interne de leur choix a posteriori. Nos résultats soulignent cependant que cet accès reflète un processus intégratif au terme duquel notre introspection assimile à la fois des données internes et des informations exogènes. Les illusions introspectives sont dès lors le résultats d'une intégration déséquilibrée entre ces différents éléments. En conclusion, cette thèse offre de nouvelles perspectives ainsi que des outils méthodologiques pour l'étude de l'émergence de la conscience des décisions. Nos résultats convergent vers l'idée que la connaissance de soi est un processus d'inférence bayésien organisé hiérarchiquement et impliquant de multiples informations.

Benjamin Rebouillat: *Beyond introspective illusions, a Brain Computer Interface approach to decision awareness*, , © September 2020

RÉSUMÉ

Nous concevons d'ordinaire nos choix comme conscients et sous notre contrôle. Toutefois, de nombreuses études montrent que nos processus introspectifs sont largement illusoires. Dans notre première partie, nous proposons que l'introspection peut être conceptualisée comme un processus d'inférence hiérarchique, et nous avançons une nouvelle approche pour en étudier les mécanismes sous-jacents. A cette fin, nous employons un protocole de prise de décision dans lequel les sujets ne peuvent accéder ni à leurs informations motrices, ni à des informations de haut niveau. En outre, nous mesurons les signaux neuronaux impliqués dans la prises de décision ainsi que la conscience que les sujets ont de leurs décisions. Cette thèse se penche sur deux questions: Premièrement sous quelles conditions peut-on être conscient de ses décisions? Deuxièmement pouvons nous accéder à nos processus mentaux par l'introspection, ou cette dernière n'est elle qu'une illusion? Nos résultats suggèrent, qu'en dépit d'un sentiment de richesse subjective, nous n'avons qu'un accès partiel aux contenus de nos décisions. Si l'on peut savoir qu'une décision est imminente, son contenu échappe à la conscience. Toutefois, les sujets peuvent accéder à une représentation interne de leur choix a posteriori. Nos résultats soulignent cependant que cet accès reflète un processus intégratif au terme duquel notre introspection assimile à la fois des données internes et des informations exogènes. Les illusions introspectives sont dès lors le résultats d'une intégration déséquilibrée entre ces différents éléments. En conclusion, cette thèse offre de nouvelles perspectives ainsi que des outils méthodologiques pour l'étude de l'émergence de la conscience des décisions. Nos résultats convergent vers l'idée que la connaissance de soi est un processus d'inférence bayésien organisé hiérarchiquement et impliquant de multiples informations.

MOTS CLÉS

Sciences Cognitives, Métacognition, Illusion, Décision, Interface Cerveau Machine.

ABSTRACT

When we make a free choice, we feel conscious and in control of our decision processes. However, over the past decades, studies on introspection demonstrated that our self-knowledge faculties are crippled by illusory content. In Part i, we suggest that introspection can be framed as a hierarchically organized inference process and we proposed an innovative methodological approach to challenge this hypothesis. We used a free decision paradigm in which no high order nor low motor level processing were solicited. Further, we track in real time internal decision variables through a Brain Computer Interface (BCI), and probe both implicitly and explicitly participants' decision awareness. The present thesis investigates two main questions. First, what are the conditions for people to be aware of their impending decisions? Second, does people's introspections access genuine mental activity or are they pure retrospective illusions? Our results suggest that despite the general impression of a rich internal life, people are only partially aware of their impending decisions. If they can consciously track their upcoming decisions, they have no con scious access to those decisions' content. Yet, when recalling their recent choices, people can access internal representation of the chosen alternative. However, our results suggest that introspection has no privileged access to internal decision variables but rather stem from an integrative process involving both endogenous and exogenous cues. Introspective illusions thus reflect an imbalanced integration process, where weak and noisy internal variables are dominated by deceptive feedback. Overall, the present thesis provides new insights and methodological tools for the study of decision awareness emergence. Our results converge toward the idea that self-knowledge of decision is a hierarchically organized Bayesian inference process involving multiple cues.

KEYWORDS

Cognitive Sciences, Metacognition, Illusion, Decision, Brain Computer Interface.