The Strength of the Mind:
Essays on Consciousness and Introspection

Jorge Francisco Morales Ladrón de Guevara

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2018

ABSTRACT

The Strength of the Mind: Essays on Consciousness and Introspection

Jorge Francisco Morales Ladrón de Guevara

I defend the view that mental states have degrees of strength. Our pains are more or less intense, our mental imagery is more or less vivid, our visual perceptions are more or less striking, and our desires and thoughts are more or less gripping. Mental strength is a phenomenal magnitude shared by all conscious experiences that determines their degree of felt intensity. Mental strength, however, has been largely ignored over other aspects of mental states such as their representational contents, phenomenology, or type. Considering mental strength is crucial for illuminating philosophical discussions related to representationalism, the transparency of experiences, cognitive phenomenology, attention, and the structure and function of consciousness. I use mental strength to develop in detail a neuropsychologically plausible theory of introspection and its limits that is inspired by a signal detection theoretic model of perception. In the second half of the dissertation, I look into methodological issues concerning the neural correlates of consciousness such as controlling for performance capacity and stimulus strength, and what these methodological concerns reveal about our theories of consciousness and its function.

Contents

Figures

Acknowledgements

I have been extremely lucky to be surrounded by many caring and dedicated people throughout my graduate years. Here, I can only try to acknowledge a subset of them. I would like to express my gratitude, first of all, to my committee members—John Morrison, Christopher Peacocke, Wayne Wu, Elliot Paul, and Hakwan Lau—not only for their care and attention in reviewing this work, but also for their guidance. John helped me navigate uncountable times the many hurdles of graduate school and philosophy. I am extremely grateful to him for his support, for always being available and, above all, for his patience. He read an unreasonably large number of drafts, and his comments—I am sure—have helped me become a better philosopher. I also thank Chris for his multiple insightful remarks on my work, and for his invaluable support throughout my time at Columbia. I am also very grateful to Wayne, who carefully read several versions of my manuscripts. His charitable approach to my work, as well as his support and advice were essential for getting me through the last mile of graduate school. In my early years in the program, I took Elliot's fantastic seminar on Descartes's *Meditations*. His example as a teacher and as a scholar made a long-lasting impression on me.

I owe a special debt of gratitude to Hakwan. He graciously took me under his wing during my first year at Columbia, even though I was clearly unqualified to work in his lab. Hakwan has patiently trained me, challenged me, and worked alongside me. His influence and generous mentorship will leave a permanent mark on my philosophical and scientific work. I should also thank Hakwan and the rest of the lab members for building an environment that provided intellectual challenges and a petri dish where ingenuity, collaboration, and collegiality thrived. Most importantly, the lab was a space that nurtured friendship. Especial

thanks to the old gang at Columbia for their help training me and for being formidable colleagues: Brian Maniscalco, Dobromir Rahnev, Guillermo Solovey, Ai Koizumi, Li Yan McCurdy; and to the new gang at UCLA for our recent collaborations: Brian Odegaard, Megan Peters, and Vincent Taschereau-Dumouchel.

Others, although not in my committee, took no less interest in my education and development. I am grateful to David Rosenthal for our many conversations. Even though I do not think I will ever convince David about the importance of signal detection theory for the study of consciousness, I will always cherish our discussions from which I learned a lot. I am also in great debt with Steve Fleming, who stoically guided me through the intricacies of human neuroimaging.

My time at Columbia would not have been the same without friends and colleagues from which I learned about the mind and beyond. For going out of their way to provide helpful and nuanced comments and criticisms of my work, I am grateful to David Barack, Simon Brown, Matthew Heeney, Andrew Richmond, and, especially, to Nemira Gasiunas. Without her support, graduate school would have been orders of magnitude more difficult. My special thanks to Alison Fernandes, Christine Susienka, and, of course, Rush Stewart, for their advice and friendship. I was also fortunate to be in a city where philosophers of mind abound. Many thanks to my friends at CUNY and NYU for enriching my time here. Many others also helped me to shape my views about philosophy and life, and to navigate the convoluted world of professional philosophy, including Zack Al-Witri, Avery Archer, Matthias Birrer, Borhane Blili Hamelin, César Cabezas, Mateo Duque, Jonathan Fine, Robby Finley, Jeremy Foster, Lydia Goehr, Isabel Kaeslin, Yang Liu, Antonella Mallozzi, Laura Martin, Katharine McIntyre, Usha Nathan, Ignacio Ojea, Michael Nielsen, Mariana Noé, Kathryn Tabb, Matti

Vuorre, and Porter Williams. And last but not least, my lifelong partners in crime, Guillermo Ortiz, Gabriel Arrache, and Daniel Vázquez.

I also thank those who, beyond philosophy and cognitive science, helped me to cross the finish line: in CUSP, Lavinia Lorch and Chanda Bennett; in the Department of Philosophy, Stacey Quartaro, Asha Panduranga, Maia Bernstein, and Clayton Rains.

I would like to express my deepest gratitude to my family. I have only enjoyed unconditional love, constant support, and understanding from my mother, Lorena Ladrón de Guevara, my father, Jorge Morales, my sister, Diana Morales, and my brother-in-law, Juan Carlos González. For their warm support, I would also like to thank my wife's family: Celia, Juan Bautista, Sofía, Carolina, José Pablo, Juan Ignacio and Sofía. Thanks to Regina and Juan Pablo for bringing us so much joy.

No words suffice to express my gratitude to my wife, Jimena Monjarás. She shared all the joys—and endured the hardships—of this long journey alongside me. Her infinite love, support, and faith in me meant everything. This dissertation is dedicated to her.

To Jimena

Preface

Conscious mental states have varying degrees of strength; pains are more or less intense, mental imagery is more or less vivid, perceptions are more or less striking, and thoughts and desires are more or less gripping. What explains these degrees of strength? What is the function, if any, of mental strength? Despite its importance, these questions about mental strength have been historically overlooked over other aspects of mental states. Philosophers have emphasized propositional attitudes, representational contents, sensory phenomenology, and type. Psychologists have emphasized stimulus strength, attention, and cognitive control. In *The Strength of the Mind: Essays on Consciousness and Introspection*, I defend the view that mental states have degrees of strength. Mental strength is a phenomenal magnitude shared by all conscious experiences that determines their degree of felt intensity. Mental strength is useful for explaining a number of phenomena. For instance, I use it to develop a systematic and neuropsychologically plausible theory of introspection and its limits inspired by a signal detection theoretic model of perception. Mental strength illuminates philosophical discussions related to attention, cognitive phenomenology, representationalism, and the transparency of experiences. Mental strength also affects the structure of our conscious stream and it reveals important functions of consciousness. In the second half of the dissertation, I present work done in collaboration with Hakwan Lau were we look into methodological issues concerning the neural correlates of consciousness such as controlling for performance capacity and stimulus strength, and what these methodological concerns reveal about our theories of consciousness and its function.

In Chapter 1, "Mental Strength," I introduce the notion of 'mental strength', a phenomenal magnitude of conscious experiences that determines their felt intensity. We report this quantitative dimension of our phenomenology when we describe the intensity of pains, the intensity of visual experiences, the vivacity of mental images. Historical precedents of mental strength are found in Hume, Kant, and William James. Mental strength is distinct from attention, stimulus and psychological saliency, and representational contents. I show the main features of mental strength in the central cases of pains, and then expand the view to mental imagery, perception, thoughts, and desires. The view that emerges is that mental strength is a domain-general phenomenal property. Incorporating mental strength to our explanations of the mental has important repercussions for how we understand cognitive phenomenology, as well as the structure and the functions of consciousness.

In Chapter 2, "A Detection Theory of Introspection," I offer a novel theory of introspection. By 'introspection' I understand the process of attentively focusing on one's current conscious mental states to form judgments about them. We can introspect sensory experiences, pains, emotions, desires, and thoughts, among other mental events. Current theories fail to explain why introspecting our experiences is sometimes easy and sometimes hard. For example, they fail to explain why it is typically easy to introspect the location of an intense pain and hard to introspect the location of a dull pain. This calls into question their adequacy. The theory I advance builds on a widespread scientific approach to how we perceive external stimuli: Signal Detection Theory. According to (SDT), the reliability of our perceptions is a function of the strength of the perceptual signal that external stimuli generate in us, so that our perceptions are more reliable when the signal is strong. For example, we perceive a person in an alley more reliably when the alley is well-lit because the

perceptual signal is stronger. Similarly, according to the theory I advance, the reliability of an introspective judgment is a function of the strength of the introspective signal that our experiences generate in ourselves, so that introspective judgments are more reliable when the signal is strong. Accordingly, I call this Introspective Signal Detection Theory (iSDT). It provides a general, systematic, and neuropsychologically plausible explanation of introspection and its limitations. It also provides insight into philosophical discussions related to the nature of perception and the transparency of experiences. I also discuss potential neural substrates of iSDT and I propose a way of testing it using psychophysical methods.

In Chapter 3, "Perception, Performance, and the Neural Correlates of Consciousness," I discuss some methodological considerations regarding perceptual strength and the quest for the neural correlates of consciousness. Studying the neural correlates of conscious awareness depends on a reliable comparison between activations associated with awareness and unawareness. One particularly difficult confound to remove is task performance capacity, i.e., the difference in performance between the conditions of interest. While ideally task performance capacity should be matched across different conditions (including stimulus strength), this is difficult to achieve experimentally. However, differences in performance could theoretically be corrected for mathematically. One such proposal is found in a recent paper by Lamy, Salti, & Bar-Haim (2009), who put forward a corrective method for EEG neuroimaging. Their analysis, however, is essentially grounded in a version of High Threshold Theory of perception, which has been shown to be inferior in general to Signal Detection Theory. The results of computer simulations are presented to confirm this along a proposal for a mathematical correction method based on Signal Detection Theory that is theoretically capable of removing performance capacity confounds. The limitations of

mathematically correcting for performance capacity confounds in imaging studies and its impact for theories about consciousness are discussed. A version of this chapter, co-authored with Hakwan Lau and Jeffrey Chiang, appeared in *Neuroscience of Consciousness* in 2015.

Finally, in Chapter 4, "The Neural Correlates of Consciousness: Theories & Functions," I focus on the different predictions current views of the neural correlates of consciousness (NCC) make. The main emphasis is placed on Two-Visual-Systems Hypothesis, Local Recurrency, Higher Order and Global Workspace theories. In particular, their predictions with respect to the role of prefrontal cortex (PFC) during visual experiences are explored. Despite the apparent stark differences between conscious and unconscious perceptual processing, available evidence suggests that their neural substrates must be largely shared. This indicates that the difference in neural activity between conscious and unconscious perceptual processing is likely to be subtle and highly specialized. The current experimental evidence about the involvement of specific activity in prefrontal cortex supports the higher order neural theory of consciousness. In consequence, imaging techniques that focus only on marked differences between conscious and unconscious level of activity are likely to be insensitive to the relevant neural activity patterns that underlie conscious experiences. Finally, it follows from the discussed evidence that the functional advantages of conscious over unconscious visual perceptual processing may be more limited than commonly thought. A version of this chapter, co-authored with Hakwan Lau, will appear in *Oxford Handbook of the Philosophy of Consciousness*, edited by Uriah Kriegel, published by Oxford University Press.

Chapter 1

Mental Strength

Hume's starting point in *A Treatise of Human Nature* is that mental states have degrees of strength: "All the perceptions of the human mind resolve themselves into two distinct kinds, which I shall call IMPRESSIONS and IDEAS. The difference betwixt these consists in *the degrees of force and liveliness*, with which they strike upon the mind, and make their way into our thought or consciousness."[1] (Hume 2000 1.1.1.1; my italics) Under IMPRESSIONS he includes sensations, perceptions, passions, and emotions, while IDEAS are "faint images of these" (*ibid*) that include memories, imaginations, reasonings, and thoughts. Hume used the degree of strength of mental states as a criterion for individuating mental states by type: "[Impressions and ideas] are in general so very different, that no-one can make a scruple to rank them under distinct heads, and assign to each a peculiar name to mark the difference."[2] (*ibid*) He has been rightly criticized for using these degrees of "force and liveliness" to distinguish types of mental states.[3] But his insight that conscious states have degrees of strength, and that these degrees are a fundamental property of conscious mental states, has been wrongly

---

[1] Throughout the *Treatise* Hume refers to this distinctive property with many different terms in addition to 'force' and 'liveliness'. Among others, he uses vivacity, influence, firmness, violence, vigor, steadiness, *strength* and *intensity* (Hume 2000, see Annotations and Glossary).

[2] The criterion was meant to have a wide scope. However, it is clear that Hume thought there could be exceptions: "The common degrees of these are easily distinguished; though it is not impossible but in particular instances they may very nearly approach to each other. Thus in sleep, in a fever, in madness, or in any very violent emotions of soul, our ideas may approach to our impressions. As on the other hand it sometimes happens, that our impressions are so faint and low, that we cannot distinguish them from our ideas. But notwithstanding this near resemblance in a few instances, they are in general so very different, that no-one can make a scruple to rank them under distinct heads, and assign to each a peculiar name to mark the difference." (Hume 2000, 1.1.1.1)

[3] For example, (Stroud 1977, 28-9; Bennett 1971, 255); but see (Everson 1988; Dauer 1999) who try to make the criterion respectable.

underestimated for too long. Pains can be more or less strong, perceptions more or less striking, mental images more or less vivid, emotions more or less intense, desires and thoughts more or less gripping. These variations in otherwise radically different states are explained, or so I will argue, by variations along a single phenomenal dimension shared by all conscious states: their *mental strength*.

According to the view I will develop in this chapter, mental strength is a distinct phenomenal magnitude of individual conscious mental states. As such, the degree of strength of a mental state can be understood as its degree of phenomenal intensity. My goal here is to develop a framework for understanding mental strength, distinguish it from related, but ultimately different, phenomena, and highlight some philosophical consequences that follow from recognizing it as a fundamental aspect of conscious mental states. To achieve this, in sections 1 and 2, I analyze the representative case of pain strength and its potential neural correlates. In section 3, the view is expanded to mental imagery, perception, thoughts, and desires. In section 4, an intrinsic and a relative understanding of mental strength is discussed. In section 5, I distinguish mental strength from attention, stimulus salience, psychological salience, and representational contents. In the last section, I discuss some of the philosophical consequences of admitting mental strength into our explanations of the mental. In particular, I discuss its repercussions in debates about cognitive phenomenology, the structure of the stream of consciousness, debates about degrees of consciousness, and the function of consciousness in general.

1. Pain Strength

Imagine you wake up late for work. You stub your big toe against the bed when rushing to the shower. A sudden painful sensation invades your conscious awareness: your toe hurts. First, the pain is sharp, strong, and unpleasant. You hold your toe and while doing so you are unable to focus on anything else except the painful sensation. After a few seconds, your experience starts changing: it slowly becomes weaker while still being a sharp, unpleasant pain in your toe. As the pain decreases, your mind gets back to thinking about being late for work and you resume your plan to take a shower.

An obvious phenomenal change takes place as the intensity of your pain first increases and then decreases. This phenomenal change is, I argue, a variation in the degree of mental strength of your painful experience.

Mental strength is a phenomenal magnitude present in all conscious experiences. It is, in this sense, domain-general. Other phenomenal properties are domain-specific. Only visual experiences have color or brightness phenomenology, only auditory experiences have loudness or pitch phenomenology, only haptic experiences have texture phenomenology, and so on. Mental strength, in contrast, is a domain-general phenomenal magnitude present, in the same way, across sensory modalities and cognitive domains. Mental strength increases from zero, as it were, when the conscious experience has not arisen, and grows in certain time to a given measure (e.g., an intense pain—stronger than yesterday's headache but milder than tomorrow's toothache). Different degrees of mental strength result in different degrees with which mental events "make their way to our consciousness," to use Hume's phrase. In other words, the degree of mental strength of a conscious state determines how intensely it is experienced, "how much it takes up" of someone's stream of consciousness at a given time.

For instance, the pain becomes "blinding" when you stub your toe. Its intensity makes it "take over" your stream of consciousness by reducing or even inhibiting the mental strength of other experiences such as your intention of taking a shower or your worry of being late for work.[4]

Pains are complex. They have distinct sensory-discriminative, affective-motivational, and cognitive-evaluative components (Langland-Hassan 2017; Hardcastle 1999; Grahek 2007). All of these components admit degrees and, together, they affect the overall phenomenology of pain. In the stubbed toe example, the affective-emotional and the cognitive-evaluative components—as well as some of the sensory-discriminative components such as felt location and sensory character (pain type)—remain constant throughout the described phenomenal changes. In contrast, its sensory intensity, which is part of the sensory-discriminative component, raises quickly and then slowly starts decreasing. In this case, mental strength changes are driven by variations along the sensory dimension.

The everyday example of hitting your toe reveals important phenomenal aspects of pains along the sensory dimension that go beyond sheer intensity. Pains have felt locations, that is, they are always felt *somewhere* in the body.[5] Phantom limb pain patients, who still feel

---

[4] Interestingly, Kant seemed to have held a similar view regarding conscious intensive magnitudes in both the *Critique of Pure Reason* ('The anticipation of perception') and in his *Lectures on Metaphysics*: "For example, when a representation has inhibited many others, we say that this has made a great impression." (cited in Longuenesse 1998, 320) Longuenesse's commentary of this passage is illuminating: "Even states of consciousness can thus be […] compared as to their magnitude. A representation is 'more or less intense' according to the multiplicity of representations it inhibits; a very great pain makes one deaf and blind toward any other representation." (Longuenesse 1998, 320)

[5] Arguably, emotional and social pains lack bodily locations. Social rejection, breaking up with a romantic partner, a relative's death, or public humiliation hurt to varying extents. However, they are not felt as such *on* the body. Certainly, it is common that bodily sensations accompany emotional pain. For example, when learning of someone's unexpected death, you may experience an empty stomach, a running heart, or a dizzy head. However, I would say that pain in these cases is not bodily, in spite of the accompanying bodily sensations. In other words, the sensations in your stomach, heart, or head are not your emotional pain, although they might hurt due to the emotional distress you are

their recent lost limb, attribute the source of their pain to a bodily location on the nonexistent limb. Pains can be felt as affecting a volumetric area or just a surface, inside or outside the body, with a precise or an undefined shape. Pains also have pain-specific phenomenal characters that determine their type; a pain can be sharp, pricking, stabbing, gnawing, burning, dull, throbbing, etc.[6]

Although mental strength naturally latches onto sensory intensity, mental strength is not exhausted by it. Mental strength can be affected by changes in other aspects of the sensory-discriminatory dimension such as felt location and sensory character, and also along the affective and cognitive dimensions, independently from changes in sensory intensity. For example, even assuming equal sensory intensities, a sharp and pounding pain may raise the overall mental strength of the experience—"how much it takes up" of the stream of consciousness—more than a dull and flickering pain. Similarly, assuming equal sensory intensities, a pain in the face may have more mental strength than a pain in the leg. For instance, the pain in the face may be more "blinding" than the pain in the leg; it may occupy more of the conscious stream of the subject, making it more distracting or attention grabbing. In section 5, I discuss again cases like this one and argue that although salience, attention, and mental strength are related, they are ultimately distinct.

The mental strength of pains can be affected too by changes in their affective-motivational dimension. The overall mental strength of an unpleasant pain may be higher

---

undergoing (pace Prinz 2005). A recent study using multivariate pattern analysis concluded, perhaps unsurprisingly but against previous findings (Kross et al. 2011), that despite their similar phenomenology, social and bodily pains are encoded differently in the brain (Woo et al. 2014). Importantly, it is clear that in both cases there are degrees of intensity.

[6] For a thorough list of sensory, affective, and cognitive aspects of pain and pain intensity used in clinical contexts, see the McGill Pain Questionnaire (Melzack 1975).

than that of an equally intense (sensorially speaking) but less unpleasant pain. For example, you may find a paper cut more unpleasant than a prick, even if you rate them as being equally intense. The paper cut may grab more your attention, affect more your capacity to focus on other things, or, in extreme cases, even "blind" you from other experiences. This suggests mental strength increases can be modulated by variations in unpleasantness too.

It may be that the sensory-discriminative intensity dimension of pain—and therefore mental strength—is increased by the changes involved in variations of unpleasantness or whether subjects rate their unpleasant pains as being more intense because of non-sensory increases in mental strength. If the former, unpleasantness is a means by which sensory pain intensity is modulated; if the latter, mental strength is a phenomenal magnitude over and above intensity in the sensory-discriminatory dimension. Although I am inclined towards the second option, for our current purposes we do not need to solve this issue and it suffices to point out these possibilities.

Someone may object that disentangling the sensory intensity and the unpleasantness dimensions is hard. Stronger pains tend to be nastier and nasty pains tend to be stronger. But sensory intensity and unpleasantness can be dissociated (Rainville 2002), which indicates they are independent from one another. For example, patients with pain asymbolia report feeling the sensory intensity of being pricked in very similar ways to the normal population. They can detect when more pressure is being exerted on them, discriminating correctly the stimulus intensity, but they do not report feeling the pain's unpleasantness (Grahek 2007).[7] This does not mean they find pain pleasant, rather they are just indifferent

[7] Some patients tended to underrate pain intensity, but they still made no adverse comments regarding the experience. In some cases, they willingly offer their hands, smiled or even laughed at the situation,

to it and they do not feel an urge to avoid it. Something similar happens after administrating morphine, thalamus lesions, or prefrontal lobotomies: patients detect the intensity of noxious stimuli in a consistent fashion, without reporting any of the suffering typically associated with them. Opiates act as if blunting the subjective appreciation of pain: "Patients who have been treated with morphine because of severe post-operative discomfort or extreme pain from cancer frequently tell their doctors, 'It's a funny thing. The pain is still there, but it doesn't bother me'." (S. H. Snyder 1996, 44) Similar, if more modest, effects are found in subjects under hypnosis (Rainville et al. 1999) and in mindful meditators (Gard et al. 2012). Importantly, the dissociation works in the other direction as well. Dental patients whose nerves are electrically stimulated while under the potent analgesic fentanyl report pain to be as unpleasant as without the drug but less intense (Gracely, Dubner, and McGrath 1979).

With respect to changes in cognitive appraisal, catastrophizers are an illustrative case. Catastrophizing, an exaggerated negative mental set brought to bear during painful experiences (Sullivan et al. 2001), affects mental strength by changing the cognitive-evaluative dimension of pain. Catastrophizing is comprised of a threefold dimension that includes magnification ("I worry that something serious may happen"), rumination ("I can't stop thinking about how much it hurts"), and helplessness ("It's awful and I feel that it overwhelms me"). Catastrophizers rate pains—usually chronic pains—as having higher intensity than non-catastrophizers with similar ailments. In contrast, when pain is reappraised and subjects stop conceiving it as a signal of a potential life-threatening pathology, intensity ratings decrease (Leeuw et al. 2007). As with unpleasantness, this opens

---

and they did not show normal physical or emotional signs such as grimacing, anxiety, or anger (Grahek 2007, 43-4).

two possibilities. One, that catastrophizing is a mechanism by which sensory intensity can be modulated; the other, that catastrophizing affects mental strength independently of the pure sensory-discriminative intensity dimension.

Mental strength determines the overall intensity of our experiences. Pain strength has multiple sources: sensory, affective, and cognitive components. This entails that some mild but nasty pains may have a high degree of mental strength; a pain that is not very strong or unpleasant, but of which we ruminate and obsess about, may have a high degree of mental strength. The interaction between these components can be complex and the mental strength of a given state need not be a simple aggregate of the contribution of each component. This suggests mental strength is not simply reducible to the specific sensory, affective, and cognitive appraisal components that modulate it. This point is important for generalizing mental strength to conscious experiences other than pain, since other mental states do not have these dimensions and they have specific dimensions that pains lack. In the next section, I discuss whether mental strength can be reduced to other general mental phenomena.

2. The Neural Correlates of Mental Strength

Pain intensity is an intuitive and phenomenally accessible way of grasping mental strength. Additionally, we can also speculate about its neural implementation. This can supplement the theory of mental strength with an extra layer of plausibility. As a property of conscious experiences, mental strength's what-it-is-like aspect shares all the well-known problems of studying consciousness and its neural correlates (Chalmers 2000; Noë and Thompson 2004;

Hohwy 2009, see Chapter 4). However, even if we remain neutral about the neural correlates of consciousness in general, we are not in the dark regarding potential implementations of mental strength.

In perception, it is fairly common to interpret measures of neural activity (e.g., neural spike rates, blood oxygen level dependent (BOLD) activity, etc.) as an indication of the strength of the perceptual response to external stimulation (Shadlen and Kiani 2013). In principle, we can subject mental strength to similar neuroscientific standards. The simplest option is that mental strength correlates linearly with neural activity related to the intensity of conscious experiences.[8] Regardless of the processes involved in rendering a mental state conscious, neural activity that correlates linearly with stimulation intensity and reports of intensity can provide clues of where and how the brain implements mental strength.

Multiple studies in fact show that in normal conditions, subjects' reports of pain intensity correlate linearly with stimulus intensity (e.g., hotter stimuli produce reports of higher pain intensity) (Coghill et al. 1999; Coghill, McHaffie, and Yen 2003; Wager et al. 2013; Atlas et al. 2014). These studies, where neural activity is measured indirectly by functional magnetic resonance imaging (fMRI) and positron emission tomography (PET), also converge in that activity in areas like insula, dorsal anterior cingulate cortex (ACC), thalamus, and somatosensory cortex shows a linear correlation with the temperature of the stimuli (and in consequence with reports of pain intensity).

These results already are indicative of a potential correlate of pain strength. But the evidence is more specific than this. Coghill and colleagues (Coghill, McHaffie, and Yen 2003)

[8] There could be nonlinear relations too, as well as patterns of activity or distinct types of computation such as divisive normalization. It is also possible that mental strength in different sensory modalities may be implemented differently from each other.

split subjects in low and high sensitivity groups by their pain intensity ratings of a 49°C stimulus. They found that BOLD-activity in somatosensory cortex and posterior ACC closely tracked the difference between low and high sensitivity groups. It is reasonable to assume from previous research that subjects are good at detecting differences in temperature, so we can confidently rule out that subjects were wrong about the temperature of the patch (i.e., we can trust that they would have reported less pain with a 48°C patch and more pain with a 50°C patch). The difference between the two groups, then, probably was how intense the pain felt to them. As I have been arguing, these differences in felt intensity are due to differences in mental strength. This entails that the difference in activity in ACC and somatosensory areas between low and high sensitivity groups is a good candidate for the neural correlates of the difference in mental strength that these subjects experienced. Note that no activity differences between these groups were found in the thalamus and insula. In other words, the thalamus and insula may encode degrees of painful stimulation, while ACC and somatosensory areas may encode the degrees of mental strength.

Whether activity in ACC and somatosensory areas are part of the neural correlates of mental strength remains highly speculative (see Legrain et al. 2011 for a thorough review). These results, however, are a proof of concept that mental strength can be subjected not only to phenomenal description but also to empirical scrutiny. Moreover, while the theory of mental strength is pitched at a phenomenological level, having a concrete proposal of its neural basis puts the theory in plausible neuroscientific terms. This will become more important in Chapter 2, where the signaling role of mental strength is argued to play a crucial role modulating the introspective accuracy.

3. Beyond Pains

So far, I have focused on pains. Mental strength, however, is a property of all conscious mental states. For brevity, here I only discuss mental imagery, perceptions, thoughts, and desires, but similar arguments can be made about other types of state such as memories, feelings, or emotions.

I argued above that pain strength has various sources: sensory intensity, unpleasantness, and a cognitive-emotional dimension. Variation in one or all of them affects the overall mental strength of a painful state, in its turn affecting how much a given painful experience "takes up" of someone's stream of consciousness at any given moment. This explanation can be naturally extended to other states. The sources of mental strength of non-painful experiences are diverse too and they interact with each other to modulate mental strength.

a. Mental imagery

The term 'vividness' is often used in the imagery domain to refer to what I mean by mental strength (Cornoldi et al. 1991; Galton 1880; Hume 2000; James 1950; Kosslyn 1996; Marks 1973; McGinn 2004; Pearson et al. 2015). Independently of their visualizing capacity, people consistently use a small set of terms to describe the strength of their mental images.[9] William

[9] At least since Galton (Galton 1880) it has been observed that the individual differences in visual imagery capacity are notable. Some people with aphantasia do not seem to have the capacity to summon mental images voluntarily at all (Zeman, Dewar, and Sala 2015).

James (1950 XVIII, vol. II), for instance, offers a compilation of reports by other scientists and his own students who, besides using descriptors like 'strength' or 'vigor', appealed to the degree of vagueness, blurriness, sharpness, dimness, clarity, or number of details to capture imagery vividness. Furthermore, subjects have no trouble providing consistent vividness ratings across time (Cui et al. 2007). This convergence strongly suggests that, even though a precise definition of vividness offers some challenges (Kind 2017), people have an intuitive understanding of the notion of imagery strength.

The strength of mental imagery is affected by, and varies along, at least six dimensions: (1) sensory properties (e.g., brightness, loudness, etc.), (2) clarity, (3) number and (4) salience of details, (5) the feeling of presence of the imagined objects or events, and (6) the overall stability of the image (Cornoldi et al. 1991; Thomas 2009). When you imagine your childhood's house, the saturation of the colors is likely to play a role in the strength of the image. The stability of the imagined house is important too. For instance, the shape of the windows may shift as you struggle to maintain the image in your mind. These shifts are clearly representational in nature. But there are non-representational components that affect the intensity of the imagined house too. The intensity of the feeling of presence that you are in front of your imagined childhood's house is unlikely to be fully representational in nature. In this case, the emotional affect attached to the image or a feeling of familiarity may play an important role too.

Naturally, all these dimensions admit variations in degree. *Ceteris paribus*, a more stable image is stronger, a brighter and more saturated image is stronger, a more detailed image is stronger, and so on. Some dimensions covary with others. Increasing the number of details can make easier imagining the objects to be present, but it might also hinder the stability of the image. Importantly, the overall strength of the image is a function of the

degrees along each of these dimensions. Mental images with faint colors and few details may still be strong. Their strength may stem from a very salient feature or from being able to picture it in a very stable way. Even within a single dimension different objects can have different degrees of strength. For example, the image of your childhood's house could be faint, but the grass and sky could be quite bright and saturated, making the image strong overall. Alternatively, even if an image is really clear or has lots of stability, few imagined details or no salient features could yield the image weak overall.

That mental imagery strength cannot be reduced to a single dimension, or to a single aspect within one dimension, can be shown as follows. If sensory properties were picked as the only dimension that mattered for imagery strength, one should not be able to strongly imagine a dim candle light. However, at least some people can do that, so imagined brightness cannot be identified with mental imagery strength. If clarity were selected as the single most important dimension, one should fail to strongly imagine a blurry image of one's childhood's house. But at least some people can strongly imagine their houses even if these have an ill-defined contour. The same applies for the remaining dimensions.

b. Perception

Perceptual and imagery strength work in a very similar fashion. Hume described imagination as perception that is "faint and languid, and cannot without difficulty be preserv'd by the mind steady and uniform for any considerable time." (Hume 2000, 1.1.3) In fact, neuropsychological and physiological research shows visual mental imagery shares many of the behavioral and neural profiles of visual perception (Pearson, Rademaker, and Tong 2011; Laeng and Sulutvedt 2014). It is not surprising, then, that this overlap has pushed

philosophers and psychologists, very much in Hume's spirit, to characterize mental imagery as perception that is "weak" (Pearson et al. 2015), "degraded" (Byrne 2010), "essentially poor" (Sartre 2004), or "decayed" (Hobbes 1962).

An important difference is that perception is committal about a particular (the representation is caused by a singular object with the attributed properties), while imaginings are noncommittal (Burge 2010, 74-5).[10] To have a perceptual experience of your childhood's house there must be a commitment regarding its presence in the immediate environment of the perceiver (this is true too in cases of inaccurate representations, illusions, and hallucinations). Imagining does not come with such commitment. Even when the imagined-to-be-present dimension is heightened, in normal cases we are still not committed to the presence of the object. Naturally, as Hume noted, cross-overs in extreme cases are not impossible (e.g., fever, madness, low threshold stimuli) (Hume 2000, 1.1.1.1). Notwithstanding this obvious difference, the other five dimensions of imagery strength function similarly in perception.

The causal origin of perceptual experiences is not *per se* relevant for our current purposes, but unlike imagery and very much like pain, perceptual strength is modulated by external stimulation. Retinal size, speed, brightness and saturation are important bottom-up modulators of perceptual strength. *Ceteris paribus*, strong stimuli give rise to strong experiences. But things are rarely *ceteris paribus*. Not just because we can misrepresent how things are, but because perceptual strength can be modulated by attention too.[11] As has been

---

[10] There are other differences too. Mental images, for instance, represent objects that need not be clearly located in space. Besides, the imagined objects exhibit an "essential poverty" to use Sartre's phrase (2004, 9). They are deprived of relations that abound in perceived objects; imagined objects are, in this sense, isolated.

[11] It is an ongoing discussion whether attention influence on appearances implies lack of accuracy or

shown in multiple experiments, attention alters appearance (Carrasco, Ling, and Read 2004; Gobell and Carrasco 2005; Fuller and Carrasco 2006; Fuller, Park, and Carrasco 2009; Montagna and Carrasco 2006; Anton-Erxleben, Henrich, and Treue 2007; Liu, Fuller, and Carrasco 2006; Tse 2005).

More could be said about the relation of representational contents, salience, attention, and mental strength in perception. However, because of its similarity to mental imagery, for the present purposes this must suffice. (See section 5 for further discussion).

c. Thoughts and desires

The case for mental strength in conscious intentional states like thoughts and desires may first appear to be more controversial. However, thoughts and desires can also dominate one's stream of consciousness over other experiences. Attention and the intentional contents of occurrent thoughts and desires may play some role in fixing the degree of strength, but they are unlikely to be the sole drivers of mental strength. A standing belief that something is highly probable, such as the believe that the sun will rise tomorrow with $0.9\overline{9}$ probability, typically has little psychological impact in and of itself. Actively thinking about this does not make an impressive dent in our conscious stream either. In contrast, thinking that there is a negligible, but non-zero, probability that one left the stove on after leaving home, often has intense psychological effects. An increase in mental strength explains why an occurrent

not (Stazicker 2011; Block 2010).

thought like this feels much more intensely than thinking about something that is more probable but that is psychologically unimportant.

Occurrent desires can be gripping too, regardless of their contents. Like thoughts, it is not the strength assigned to the propositional attitude what determines their mental strength (e.g., «I desire that *p* to degree *x*»). The well-being of my relatives is something I strongly desire, but it is a desire that hardly ever dominates my stream of consciousness. At least not in the same way a strong pain or thinking that I left the stove on dominate it. In comparison, my current desire of standing up away from the computer for a few minutes is, all things considered, less strong than the well-being of my relatives. To paraphrase Hume, I would not prefer standing up over saving the whole world. The degree of desire assigned to the propositional attitude in an all-things-considered sense is much lower than my desire for the well-being of my relatives. However, my desire to stand up is currently much stronger in terms of its mental strength; it plays a much stronger role in my current psychology and it is more central to my conscious awareness. At least right now, my desire to stand up has more mental strength than my more noble desires.

d. Domain generality

Mental strength is a phenomenal magnitude present in all kinds of conscious states. This fact is evidenced when increasing mental strength in one experience reduces the mental strength in others. When you stub your toe, your worries about being late for work and your thoughts of taking a shower disappear as pain becomes more prominent. It is almost as if the total distributed mental strength one can have at a given moment was capped and therefore shared among simultaneous states in one's stream of consciousness. This could explain why

mental states with high strength become "blinding". It is because they dominate the stream of consciousness in detriment of other states. For instance, an effective remedy against pain consists in increasing the strength of other experiences. This could be achieved by focusing attention on something else, as well as by thinking about other things, imagining one is in a different situation and, in the most extreme cases, by inducing a new source of pain. Something similar happens while vividly daydreaming. A faint image that flickers in and out of consciousness is not very distracting. But when someone gets completely lost daydreaming or meditating, imagination can be so strong that one becomes perceptually "decoupled" or unaware of one's surroundings (Hove et al. 2016; Schooler et al. 2011).

Research in psychophysics shows something like blinding takes place within the perceptual domain too. Lavie's Load Theory holds that perception is automatic and it has a limited capacity. It predicts that perceptual load in a task modulates whether other stimuli enter conscious awareness or not independently of attention and the properties of non-target stimuli (Lavie, Beck, and Konstantinou 2014; Forster and Lavie 2016). Here, perceptual load is understood as the number of target items that need to be perceived in a task. For example, in a search task, when perceptual load is high, conscious awareness includes only the main task's stimuli and it does not "spillover" to others. When perceptual load is low, other stimuli enter awareness too, distractors affect task performance, attentional capture is increased, etc. (see Lavie, Beck, and Konstantinou 2014 for a review). When the target is very different from the non-targets or when there are few non-targets, subjects become aware of task-irrelevant stimuli. These effects are found across identical attentional conditions and independently of whether stimuli are in the periphery or at fixation, whether they are objects of socio-biological significance or not, or whether subjects are expecting the task-irrelevant

stimuli or not. In contrast, when perceptual load is high, task-irrelevant stimuli go unnoticed, as if the main task blinded subjects from seeing them.

While Lavie's theory is cashed out in terms of perceptual processing and informational load, its results could be recast in terms of mental strength. As I argued above, mental strength is increased by quantitative stimulus properties, number of features, and so on. In my terms, then, Lavie's results confirm that when perceptual experiences are strong, other experiences (or other objects in the visual experience) become weaker.

Like sensory experiences, intense desires and thoughts can also become "blinding." Obsessive impulses and obsessive thinking can become distracting to the point of eliminating other states from someone's stream of consciousness.

Incidentally, the phenomenon I have labelled as "blinding" goes against Hume's goal of using mental strength to rank *types* of states. Hume's idea of strength as a means of layering types of mental states by the range of mental strength they normally have does survive scrutiny. Imaginations and thoughts, which are conceived to be systematically weak in Hume's ontology, can dominate over allegedly stronger states such as perceptions or pains.

One of the lessons of this section is that "blinding" is possible because the degrees of strength of simultaneous states interact with each other. In the next section, I explain how mental strength, which is an intrinsic property of conscious experiences, can explain "blinding" in particular, and the structuring of the stream of consciousness in general, when it is considered in a relational way.

4. Intrinsic and Relational Mental Strength

Mental strength is an *intrinsic* magnitude of phenomenal states that determines how intensely they are experienced.[12] Derivatively, mental strength is the measure of phenomenal prominence of a given conscious state at a given time in someone's stream of consciousness. This means that mental strength effectively structures the stream of consciousness. This structuring role allows mental strength to be understood in a *relational* way.

The following analogy may help clarify the intrinsic and relational aspects of mental strength. Consider New York City's skyline. The height of each building is completely independent of the heights of other buildings; it is one of its intrinsic features. The building's height affects other things: how long it takes to climb its stairs, the length of its shadow at a particular time of the day, its weight, etc. However, height in itself is not sufficient for determining the prominence of the building within the skyline. Prominence is determined by the building's relative height compared to other buildings in the city. At some point in the early twentieth century, the 22-stories of the Flat Iron Building made it stand out above the rest. It was the most prominent construction of the Manhattan skyline. However, even though the height of the Flat Iron Building has not changed since its completion

[12] It is hard to speculate about the precise nature of mental strength as a magnitude. However, insofar as mental strength is a property of subjective experiences, it should be thought of as an intensive magnitude (e.g. like temperature), rather than an extensive magnitude (e.g. length). Intuitively, an intensive magnitude allows for rankings of magnitudes of the same type (e.g. greater than, lesser than, equal to). For example, we can say that one object is hotter or cooler than another one. We can certainly rank conscious experiences by their strength: this toothache is stronger than this headache. However, intensive magnitudes do not allow assignments of a ratio to two unequal magnitudes. It is not meaningful to say that a cup of tea is twice as hot as a glass of cold water. In contrast, it is meaningful to talk about ratios of extensive magnitudes, for which an intuitive notion of addition of the magnitudes has application (Peacocke 2015). An object can have twice or half the length of another object. But, can we say that this toothache is twice as strong as this headache? Probably not.

(intrinsically, its height has remained constant), its prominence within the skyline has completely eroded due to the tens or even hundreds of taller buildings that have appeared since it was constructed (relationally, its height has changed).

Similarly, the mental strength of a conscious state is intrinsic when considered on its own, and relational when considering the structuring role it exerts upon the stream of consciousness. A mild headache may not be equally prominent when the rest of your other conscious states are also weak than when you have some other strong experience. For example, a mild headache may be prominent in the totality of your conscious stream when you are quietly reading at home. In contrast, an identical headache might not be prominent at all if you are having a lively conversation with your friends. According to the view I have been putting forward, this scenario is explained by the higher mental strength of the conversation's experience effectively reducing the prominence of your mild headache. Just as the Flat Iron Building lost prominence throughout the twentieth century, your headache loses prominence in your stream of consciousness throughout the conversation with your friends. Thus, it is the relative strengths of all your simultaneous states what determines their influence in the stream of consciousness.

This dual role of the intensity of experiences finds echo in what philosophers such as William James and David Hume himself have said about consciousness and mental strength. James thinks that the stream of consciousness, and thereby how we experience each individual state in it at any given moment, is largely dependent on other states. While discussing perceptual experiences, James writes: "What appeals to our attention far more than the absolute quality or quantity of a given sensation is its *ratio* to whatever other sensations we may have at the same time. […] We feel things differently according as we are sleepy or awake, hungry or full, fresh or tired […]." (James 1950 232-3, IX, vol. I) These

differences, however, never make us "doubt that our feelings reveal the same world, with the same sensible qualities and the same sensible things occupying it." (James 1950, 233, IX, vol. I) Here, James's suggestion is that we can distinguish the contents of our perceptions from both the absolute and relative strength with which they impact our minds at any given moment.

James thinks that the importance of strength proportions in shaping our experiences applies to thought too, not just perception. When entertaining the thought «The pack of cards is on the table», it is reasonable to say that we are entirely conscious of it throughout the duration of the thought. However, James points out, we are not equally conscious of each part as the thought progresses in our stream of consciousness. In the first, say, half second, the words 'the pack' are more prominent in consciousness; in the next half second, the words 'of cards' will gain prominence; and in the end, the words 'is on the table' dominate over the rest of the sentence. Moreover, according to James, mental strength and how it shapes our stream of consciousness is reflected in its neural correlates:

> [W]e may be sure that, could we see into the brain, we should find the same processes active through the entire sentence in different degrees, each one in turn becoming maximally excited and then yielding the momentary verbal 'kernel', to the thought's content, at other times being only sub-excited, and then combining with the other sub-excited processes to give the overtone or fringe. (James 1950, 282, XI, vol. I)

David Hume's own view is also consistent with the intrinsic and the relative aspects of mental strength I have highlighed. The terms 'force' and 'liveliness' used by Hume to describe mental strength, "refer to intrinsic properties of images and are non-relational. Whether, and to what degree, a 'perception' has force and vivacity can be determined by examining that perception *by itself*." (Everson 1988, 404) On the other hand, "Hume is explaining force and vivacity in

functional terms. One 'perception' has greater force or vivacity than another if it is such as to produce a stronger effect on the mind." (Everson 1988, 406) In Hume's own words, mental strength causes experiences "to weigh more in the thought, and gives them a superior influence on the passions and imagination." (Hume 2000, 1.3.7.7) Moreover, Hume also thinks that we can distinguish between the mental strength of a conscious mental state and its contents (see next section). Impressions and ideas in general, and ideas of memory and ideas of imagination in particular, differ "in having a greater degree of vivacity, or force and liveliness-conceived not as an additional perception or mental content but rather as a 'manner' in which these ideas occur" (Garrett 2002, 26). Hume refers to this 'manner' also as a 'feeling' that varies in its degree of intensity (Hume 2000, 1.3.7.7).

I will have more to say about the structuring functions of mental strength in the last section. Now, I move on to discuss potential alternative explanations for the phenomena whose underlying cause I have been attributing to mental strength.

5. What Mental Strength Is Not

One may wonder, from the examples in the previous sections, if mental strength is truly a distinct trait of mental states. After all, the cases I discussed seemed to involve either bottom-up stimulus driven intensity changes or top-down attention effects. Mental strength could just be garden-variety sensory intensity driven by the representational contents of the state involved (e.g., potential tissue damage, a bright light, etc.) that goes up when attention is involved and goes down when it is not. For states like desires or thoughts, mental strength

might really just be the degree of occupation of attention or cognitive access. If this were so, mental strength would not be a distinct, domain-general phenomenal magnitude of conscious experiences. It would be, at best, a domain-specific phenomenal magnitude. At worst, mental strength would be reducible to something else. I identify attention, stimulus salience, psychological salience, and representational contents as the most likely suspects. I will argue that mental strength is not reducible to either of them. For simplicity, I will limit my arguments to sensory states such as pains, mental imagery, and perception. However, I think the distinctions I lay out in this section could be extended to intentional states like desires and thoughts too.

a. Mental strength is not attention

A potential objection to the distinctness of mental strength is that it is just attention.[13] *Prima facie*, mental strength and attention covary with each other. Moreover, it is well known that attention affects phenomenology (Carrasco, Ling, and Read 2004), and that it can be captured by sudden stimuli (so-called exogenous attention) (Wright and Ward 2008). One could argue that what I call mental strength really is the orientation of attention.

Despite the initial plausibility of this objection, attention and mental strength are distinct. First, they are not in the same metaphysical category: mental strength is a property of conscious experiences themselves, whereas attending is an activity of a cognitive capacity.

---

[13] It is not a coincidence that similar notions to mental strength, but ultimately different from it, involve attention. Some of such notions include Beck and Schneider's mental primer (2017) and Wu's phenomenal saliency (2011).

Second, while mental strength can be *modulated* by attention, they are not identical. Presumably, your toe pain comes into existence *before* you attend to it. Or at least, counterfactually, we can say that attention would not have been directed to your toe had you not experienced pain. It would be odd to claim that the cause of your pain or the cause of the intensity of your pain is *that* you attended it. The order of explanation seems backwards.[14]

This observation does not annul the important role attention performs in modulating strength. In the case of pains, for example, subjects whose attention is distracted away from a noxious stimulus (e.g., by engaging cognitive resources in a demanding task), generally report less intense pains (Miron, Duncan, and Bushnell 1989; Legrain et al. 2009; Bantick et al. 2002). The opposite effect takes place too. When noxious stimulation becomes the main focus of attention, subjects rate pains as being stronger (Miron, Duncan, and Bushnell 1989), and physiological markers and neural activity of areas known to code for pain strength become more active (Hauck, Lorenz, and Engel 2007).[15] It is remarkable that even in cases where external stimulation is lacking altogether (e.g., in phantom limb patients), pain strength is also modulated by attentional factors (Nikolajsen and Jensen 2001). However, and to the point of its distinctness, attention typically modulates pain strength only within a limited range. Normally, even if you give a weak pain your full attention, it does not become

[14] Some philosophers and psychologists have argued attention is necessary for consciousness (M. A. Cohen et al. 2012; Prinz 2012). They would probably disagree that a conscious pain is felt before being attended. The debate is complex. Here I only would say that my view is compatible with unconscious attention, or subpersonal mechanisms necessary for attention, being engaged before pains become conscious. They may even cause it to become conscious. However, I would maintain that subject level, conscious attention is attracted to the pain only as a result of its being conscious in the first place.

[15] This kind of reverse inference, however, requires careful consideration (Poldrack 2006; Machery 2014).

excruciating. Alternatively, you can try to distract your attention away from an excruciating pain and, normally, it will remain quite strong.[16]

The case for distinguishing mental imagery and attention is similar. Attention increases the strength of mental images. By focusing on the generated image and attending its features, more clear, stable, salient, and bright details may be experienced, plausibly making the image stronger. But this kind of attentive focusing need not translate into strength at all. First, increases in one isolated dimension do not necessarily imply increases in the strength of the image overall. Second, sometimes we just fail, in spite of our efforts, to picture a strong image. Thinking otherwise would amount to saying it is always in our hands to generate strong images given that it is always in our hands to attend to their features. Rather, like in the case of pains, attending can enhance the strength of an already existing image only within a limited range. And the same reasoning applies to perception.

b. Mental strength is not salience

Salience is another candidate one might feel tempted to invoke when discussing mental strength. In vision science, it is well established that salient stimuli grab attention (for a review see Itti and Koch 2001), they alter appearance by increasing apparent contrast and

[16] Someone may argue that sometimes there are paradoxical effects of directing attention to pains. For example, during mindfulness therapy, subjects report experiencing reduced pain despite the fact that the mindfulness method involves attending to different aspects of the experienced pain (called 'sensory splitting'). However, mindfulness also requires focusing on other mental states, activating mental imagery of pleasant scenery, and letting thoughts and sensations simply pass. Moreover, in experimental settings where pain reduction has been found using mindfulness, participants typically do not meditate *while* painful stimulation is being delivered but a few minutes beforehand (Zeidan et al. 2015; Zeidan et al. 2010). Overall, attention during mindfulness is more distracted away from the pain than it is focused on it.

apparent saturation (Kerzel et al. 2011), they increase performance and reaction time in a wide variety of tasks (Donk and van Zoest 2008) and, in general, salient stimuli effortlessly stand out from their neighbors in a visual scene. A salient stimulus, say, a red letter in a page full of black letters, immediately seems to attract attention, it facilitates detection during a visual search, and, one might argue, it increases the degree of mental strength of the stimulus. Pain strength may also be attributable to pain salience. After all, the sensation of pain in your toe stands out from the non-painful sensations in the neighboring parts of your body, and the effects of pain salience and pain habituation are very similar to those in visual perception (Legrain et al. 2011).

Like attention, salience can be a modulator of mental strength (J. Beck and Schneider 2017). For example, Kerzel and colleagues (2011) showed that salient stimuli (e.g., a tilted bar in a set of upright bars) increased their apparent contrast and color saturation. It is reasonable to suppose that the mental strength of the experiences of a tilted bar was heightened along these appearance changes. However, mental strength cannot be identified with salience. The most obvious reason is that salience, as it is understood in psychology, is typically a property of stimuli, not of the mental states that represent them. But even if we focused on the experiences themselves, it is obvious that they still have a degree of strength when there is nothing salient in the environment they represent. If shown a display with a homogeneous set of stimuli (say, a matrix of white dots on a black background), your experience of the individual dots will have certain mental strength, even though none of the dots is (represented as) salient.

Another reason for not identifying salience with mental strength is that the mental strength of an experiences of a salient stimulus can change depending on the mental strength and attention demands of experiences of non-salient stimuli. For example, in an EEG study,

when subjects performed an easy visual task, salient nocive stimuli produced larger event-related potentials (ERP) in the P400 component compared to homogeneous nocive stimulation. However, the difference in ERPs associated to the same set of salient and homogeneous nocive stimuli decreased when subjects simultaneously performed a visually demanding task (Legrain et al. 2005). I would not say that the P400 component is necessarily a physiological marker of pain strength. However, it is consistent with all I have said in this chapter (especially with what I said about "blinding" and the structure of consciousness) that the visual experiences become relatively stronger than salient pains when subjects performed the more demanding task. In that case, the P400 component could be a marker of relative pain strength.

Finally, mental strength and salience seem to provide different time courses and benefits. For instance, performance is increased with salient stimuli, but this benefit is only short lived (Donk and van Zoest 2008). Salient stimuli improve performance in visual tasks, but only if the response is produced very fast (in less than 200ms after stimulus onset).[17] Responses with longer latencies—which constitute the majority of responses we produce outside the lab—do not seem to benefit as much from salience. The effects of a mentally strong state, however, seem to last much longer than a few milliseconds. When a strong experience dominates the stream of consciousness, it attracts attention and rearrange the subject's mental structure for several seconds or even minutes.

[17] Responses in this study were made with saccades (rapid, ballistic eye movements). Perhaps other response modalities preserve performance benefits longer.

These results, along with the ones discussed in the previous subsection, suggest an intricate relation between mental strength on one hand, and attention and salience on the other. This relation, however, is not that of identity.[18]

c. Mental strength is not psychological salience

According to a recent proposal, attention is an "activity of creating, maintaining, and changing a certain structure of the mind" (Watzl 2017, 70). Watzl's view, according to which attention regulates mental priority structures, provides a good contrast with the view about mental strength that I am presenting here. He suggests, correctly I think, that the elements of the mind, namely, its mental states, are organized in priority structures. Priority structures are regulated by attention, and attention is guided by psychological salience (passive attention) and executive control (active attention). Unlike stimulus salience, psychological salience is a property of mental states themselves.

According to Watzl, "an occurrent state is passively attention-guiding in virtue of being psychologically salient. And it is psychologically salient, because it presents an *attention command*. When priority structures evolve passively, they follow *psychological imperatives* issued from within those structures themselves." (Watzl 2017, 115) In other words, when subjects are not actively guiding their attention, mental states themselves become like

[18] Here I did not discuss the relation between salience and attention. While some have argued for a necessary connection between conscious attention and phenomenal salience (Wu 2011), there is some empirical evidence that suggests attention and salience are independent (Kerzel et al. 2011).

"basins of attraction" for attention. For him, "psychological salience consists in having an imperatival content of roughly the form <put *x* on top of a priority structure!>." (*ibid.*)

Whether psychological states have, in addition to their regular representational contents, imperative contents to which subject-level attention is sensitive to is something I do not discuss here. I cannot help but express skepticism, although it is irrelevant for my current purposes. What matters here is that psychological salience is not what I mean by mental strength. I take mental strength to be conscious intensity, a phenomenal magnitude, which cannot be simply identified with any kind of content (see next subsection). Thus, mental strength cannot be some kind of imperative content. Moreover, Watzl thinks that priority structures also have unconscious states as parts (2017, 76, and chapter 12) and, therefore, psychological salience can be a property of unconscious states. I do not dispute that attention can be grabbed by unconscious stimuli (Jiang et al. 2006), and maybe unconscious mental states have some kind of psychological salience that attracts attention unconsciously, which, thereby, regulates an all-encompassing priority structure. Nevertheless, mental strength is exclusively a property of conscious experiences and because of that it cannot be equated to psychological salience. Finally, perhaps somewhat surprisingly, Watzl thinks that psychological salience is primarily a property of perceptual states (and perhaps other sensory and emotional states). However, as I have been arguing, mental strength is a phenomenal property of all conscious mental states, including occurrent thoughts and desires.

Watzl's theory is thorough and sophisticated, and further exploration of its commonalities and differences with mental strength is warranted. For instance, much of the work priority structures and psychological salience do in Watzl's theory is explained by mental strength. Thus, ours could be rival or complementary explanations of related

phenomena. For now, however, it must suffice to say that mental strength is not identical to psychological salience.

d. Mental strength is not representational contents

Two questions may be asked about the relation between mental strength and representational contents. First, is the intensity of experiences modulated by their representational contents? Second, is the intensity of experiences reducible to their representational contents? An affirmative answer to the first question would indicate that the representational status of experiences and their mental strength are related, but mental strength might still be a distinct phenomenal magnitude. An affirmative answer to the second question would put pressure on the claim that mental strength is distinct. Here I will argue that mental strength may be modulated by the representational contents of experiences, but that strength is ultimately distinct from representational contents. For ease of exposition, I will focus on pain strength.

Claiming that pain strength is distinct from whatever representational contents pain experiences may have (contra Armstrong 1968, 314-5; Bain 2007; Cutter and Tye 2011; Tye 1995), is not to deny the commonsense observation that external stimulation is an important modulator of pain strength.[19] In fact, like with felt unpleasantness, pain intensity and tissue

[19] This seemingly commonsense observation has a convoluted path throughout the history of philosophy. For Aristotle, pain is a nonspecific affect that accompanied every sensation (Aristotle 1994, II(3), 414b). It was Avicenna the first to seriously consider pain as a distinct sense (Dallenbach 1939). By Descartes' time, pain was fully understood to be a distinct sensory modality. Descartes thinks that "if there is some bodily damage, there is a sensation of pain" (Descartes 1985b, I:282 , AT VIIIA 318) and that "the sensation we call 'pain' always results from an action so violent that it injures the nerves" (Descartes 1985c, I:362 , AT XI 399-400). But Descartes later developed a more nuanced view about

damage often co-vary. Thermal stimulation, for instance, is reported to be more painful in a linear fashion (Dubin and Patapoutian 2010; Coghill et al. 1999) and associated distributed brain activity increases linearly as well (Coghill et al. 1999).[20] Presumably, this extends to simple cases of chemical, electrical, and mechanical pain too.

In a slogan, representationalist philosophers about the phenomenal character of conscious experiences hold that "if two experiences are alike representationally, then they are alike phenomenally (and vice versa)."[21] Hence, contrary to my proposal, the representationalist would say that changes in mental strength are really just changes in represented strength (Dretske 1995; Byrne 2001; Harman 1990; Tye 2000).[22] The pain representationalist's argument could unfold as follows. Suppose that a painful state's content is something like "«there is a disturbance of type *d* in location *b*»" (Cutter and Tye 2011, 92).

---

the relation between stimulation and pain. In the *Treatise on Man* (1985d), he develops a mechanistic account of pain according to which the amount of damaging nerve stimulation, which in turn proportionally affects the nerve firing *pattern*, determines pain intensity. Based on cases of phantom limb pain, he argues that "the cause of pain lies in the other areas through which the nerves travel in their journey from the limbs to the brain" (Descartes 1985b, I:283, AT VIIIA 320). While rudimentary, Descartes' understanding of pain is closer to contemporary theories of specificity, gate control, and coding patterns (Cervero 2012; Moayedi and Davis 2012; Prescott, Ma, and De Koninck 2014).

[20] This linear relationship has limits. External stimulation starts being painful only above a certain threshold (about 45°C for thermal pain) and it eventually stops hurting if tissue damage is such that nociceptors are completely destroyed (above 50°C nerve damage starts occurring). Furthermore, the most promising physiological theories of pain hold that the strength, sensory, and affective characters of pain are built from a series of opponent processes at the spinal and cortical levels rather than from simple one-to-one mappings between stimulation and experience (something akin to the opponent theory of color perception) (Prescott, Ma, and De Koninck 2014).

[21] This slogan marks the commitment of what has been called 'weak representationalism'. 'Strong representationalism', in contrast, holds that the qualitative character of our experience *consists in* the representational content of such states (Tye 2000). The following discussion addresses weak representationalism.

[22] Representationalism is part of a wider view called intentionalism, according to which phenomenal characters can be reduced to contents, even if not representational ones. For pains, intentionalism can take the shape of imperativism, according to which the contents of painful experiences are commands (Martínez 2010; C. Klein 2015). Imperativists have recently addressed the issue of pain strength (C. Klein and Martínez 2016), but here I only address representationalist concerns.

Assume too that "the physiological type *d* includes information about the shape, volume, and *intensity* of the disturbance" (*ibid.*; my emphasis). Thus, under disturbance type, spatial extent (i.e., shape and volume) and the intensity of pain are included.[23] The changes in felt strength would be effected by changes in how spatial extent and intensity are represented. As with any other representation, the representational accuracy of the actual spatial extent and intensity of the tissue damage can vary. Phantom limb pain, for instance, would be an extreme case of inaccurate representation. But even in that case, pain intensity could be explained as the (inaccurate) representation of (potential) tissue damage.

Let us focus on attention and representational accuracy. In the perceptual domain, attention systematically makes subjects faster and more accurate when discriminating stimuli (Posner, Snyder, and Davidson 1980; Posner 1980; Wright and Ward 2008; Carrasco 2011). In the classic Posner attention paradigm, subjects discriminate (detect or identify) a stimulus briefly presented at one of two possible locations while directing their gaze to a central fixation point throughout each trial. A cue indicating with a certain probability the location of the next target is briefly presented before stimulus onset. Subjects are instructed to use this cue to direct their attention toward the expected target location. Their responses are systematically faster and more accurate in valid/attended trials (i.e., when the target appeared at the predicted location) than in invalid/unattended trials. There is a consensus that these behavioral improvements are achieved via perceptual signal processing

[23] In earlier formulations (Tye 1995), the spatial extent of tissue damage seemed to be identified with the intensity of pain. In this more recent presentation, Cutter and Tye seem to assume these are orthogonal dimensions. For simplicity, I address them together.

enhancement and noise reduction that lead to increased representational accuracy (Carrasco 2011).[24]

If mental strength is just represented tissue damage, it is not surprising, the representationalist would argue, that attention affects the precision of the relevant pain representations. As noted in the preceding subsection, attending increases pain and distracting attention decreases it. The representationalist would say this is not surprising because this modulation of representational precision is well established for the perceptual domain and pain—that is, the representation of potential tissue damage—is not different (cf. Aydede 2009). Consider this example. Let us stipulate that the actual extent and intensity of the bodily disturbance that produces your toe pain has 5 arbitrary units (a.u.). Then, you try to ignore your pain by occupying your attention with something else. The effect, we know, will be the reduction of pain. The alleged explanation is that your pain represents inaccurately the extent and intensity of bodily disturbance when you distract your attention and, thus, you now experience, say, 3 a.u. of pain.

This explanation is consistent with the experiments described in the previous subsection. Despite its *prima facie* plausibility, however, this explanation cannot fully account for the data. As we saw above, inattention systematically *decreases* pain strength. But there is nothing about inaccurate representation due to inattention that requires *unidirectional* inaccuracy. It is hard to see what a representationalist explanation would be. Appealing to inattention does not explain why, when distracted, subjects do not feel stronger pains sometimes. Inaccuracy implies variability in any direction. Why are inattentive subjects not

[24] The precise neural mechanisms enabling this enhancement and noise reduction are the object of current research (Carrasco 2011; Carrasco et al. 2013; Desimone and Duncan 1995; M. R. Cohen and Maunsell 2009; Reynolds and Heeger 2009; Pestilli et al. 2011).

inaccurate by representing, say, 7 a.u. of pain instead of 3 a.u.? If attention increases representational accuracy, attending should decrease, not increase, pain intensity if inattentive subjects were being inaccurate by overestimating their pain. But decreasing pain intensity with attention goes against the evidence presented above.

The representationalist could try to insist that subjects are systematically biased to underestimate the extent and intensity of bodily damage when they are not attending. While not impossible, a systematic bias for being wrong in one particular direction in this case would be bizarre. It would be bizarre in the perceptual domain too. It would be surprising to discover that when not paying attention, humans always see things, say, 10° of visual angle to the left of where they really are. Note that I mean misrepresenting the location of objects, not just having a computational bias. Certainly, unidirectional computational biases in the perceptual domain are not unheard of. For example, our visual system solves convex-concave ambiguity by assuming light comes from above (and slightly to the left) (Sun and Perona 1998). However, this is not a bias that makes us systematically *wrong*. Rather, it is a computational bias that makes us, in fact, accurate on a vast majority of times *despite* informational ambiguity. Furthermore, a systematic underestimation of the extent and intensity of bodily disturbances is not a prediction of representationalism and it would seem *ad hoc* to assume it unless independent reasons were offered. In contrast, the evidence can be simply explained by appealing to a direct modulation of pain strength by attention.[25]

---

[25] An interesting case is precisely that of attention altering appearance. It has been repeatedly shown that attention alters appearances along several dimensions. Typically, these changes take place in one direction (i.e., stimuli become brighter, larger, etc.) (Carrasco, Ling, and Read 2004; Gobell and Carrasco 2005; Montagna and Carrasco 2006; Fuller and Carrasco 2006; Fuller, Park, and Carrasco 2009; Anton-Erxleben, Henrich, and Treue 2007; Liu, Fuller, and Carrasco 2006; Tse 2005). I think that this is *precisely* because what is being altered in those cases is mental strength, rather than the perceptual representations. This would be consistent with those views that hold that changes in

In summary, felt pain strength can vary independently of the representation of external stimulation. The evidence for this is largely due to experiments that manipulate pain strength via attention while keeping stimulation constant. I argued that these results cannot be easily explained by a representationalist account of pain strength. This does not mean I have refuted the representationalist position as such.[26] For example, I have not shown that pains do not have contents (they probably have some), that they do not have representational contents of the extent and intensity of external stimulation (they probably have some), or that other phenomenal properties of experiences are not reducible to representational contents. However, the objections laid out here against the representationalist position make plausible that mental strength is a distinct phenomenal property of painful experiences.

6. Further Consequences

Characterizing mental strength is a valuable project in its own right, independently of its philosophical consequences. Nevertheless, making explicit some of these is important. I will finish by pointing out how mental strength sheds light onto some relevant philosophical issues related to cognitive phenomenology, the structure of the stream of consciousness, the debate about degrees of consciousness and, finally, the functions of consciousness.

---

appearance through attention do not necessarily involve inaccurate representations (for discussion, see Stazicker 2011; Block 2010).

[26] For arguments against pain representationalism, see (Aydede 2009; Aydede 2017).

a. Cognitive phenomenology

It is not a trivial finding that mental strength is a domain-general phenomenal magnitude. Phenomenal character is often described only in terms of domain-specific sensory qualities: the redness of a tomato [vision], the sweet savor of pineapple [taste], or the odor of a skunk [smell]. Mental strength, in contrast, is present in all experiences in spite of originating from diverse phenomenal and representational components unique to each domain. An important consequence of the domain-generality of mental strength is that cognitive states such as thoughts and desires have a non-sensory phenomenal magnitude associated to it. This does not address whether cognitive states have *distinct* phenomenology from sensory states, which is the question that drives most of the debate about cognitive phenomenology (Chudnoff 2015; Bayne and Montague 2011). However, the domain-generality of mental strength indicates that cognitive states have phenomenology that is not sensory-specific. This much can be accepted even if it turns out that cognitive states lack any other type of phenomenal character.

b. The structure of the stream of consciousness

Philosophers and psychologists often describe the structuring relations between mental states using the familiar terminology of 'center' and 'periphery'.[27] Naturally, this structuring relation does not have to be binary, it could be graded. On my view, mental strength is the

[27] There is a large tradition within Gestalt psychologists and phenomenologists like Husserl, Sartre and Merleau-Ponty. For a recent approach see (Watzl 2017).

structuring property of the stream of consciousness. This should be clear from the relational reading of mental strength I discussed in section 4. In other words, when mental strength is considered in our explanations of the mental, structuring comes for free.

This is not true of all proposals that highlight the importance of central/peripheral relations. For example, Watzl's priority structures theory requires both bottom-up and top-down attention (see section 5.c above). On my view, in contrast, the explanation is straightforward and uncostly. Conscious mental states have an intrinsic property, mental strength, by means of which they fall into a natural ordering. Our conscious life *is* structured, rather than having to be constantly structured. The difference is subtle but important. In Watzl's view mental states are, so to speak, inert. It is the powerful action of the constant deployment of attention that keeps them ordered. This should not rest importance to attention as a source of mental strength and, hence, as a powerful structuring tool. In the case of voluntary attention, it is a subject-guided structuring tool, which entails that the shape of our conscious life is to a large extent under our control. This, however, does not mean that attention is the ultimate explanation of the conscious mind's structure. Rather, the structure of the conscious mind depends on mental strength. Thinking otherwise risks attributing attention powers it does not have, as it was sharply pointed out by William James:

> Thus the notion that our effort in attending is an original faculty, a force additional to the others of which brain and mind are the seat, may be an abject superstition. Attention may have to go, like many a faculty once deemed essential, like many verbal phantom, like many an idol of the tribe. It may be an excrescence on Psychology. *No need of it to drag ideas before consciousness or fix them, when we see how perfectly they drag and fix each other there.* (James 1950, 452, XI, vol. I; my emphasis)

We should agree with James that there is no need of an "additional force" to structure

consciousness other than mental states themselves. I would not go as far as to deny the existence of attention and its important role structuring the stream of consciousness. But the distinction between attention and its effects, as well as their limits, must be clear. The conscious mind is ultimately self-structuring.

c. Degrees of consciousness

There is a lively debate in philosophy, psychology and neuroscience about whether consciousness comes in degrees (Dehaene 2014; Bayne, Hohwy, and Owen 2016; Fazekas and Overgaard 2017; Rosenthal 2018). When talking about consciousness and its cognates, researchers may refer to the overall state of a person or animal (e.g., wakefulness, anesthesia, coma, sleep, etc.), for which they use phrases such as 'state-consciousness' or 'global states of consciousness'. They may also refer to whether the person or animal is conscious of something or not (e.g., seeing or not seeing a face, seeing a face versus seeing a house, hearing or not hearing a sound, feeling or not feeling pain, etc.), also referred to as 'content-consciousness' or 'local states of consciousness'.[28] The question about degrees of *local* consciousness sometimes is framed in representational terms. Whether a subject is more or less conscious *of* something is taken to be a question about whether the *representational contents* of the subject's state are more precise, more complete, or more intense.

[28] This terminology is more common in neuroscience than in philosophy, but I find it clearer. Confusingly, in philosophy the terms 'transitive-' or 'state-' consciousness are often used to refer to what I labelled here 'content-consciousness', and the term 'creature-consciousness' is sometimes used to approximate what I called here 'state-consciousness' (Rosenthal 1993).

I think that the question about degrees of local consciousness should be, or at least it can also be, interpreted to be about phenomenal states themselves. Under this reading, we can take into account the representational contents of experiences insofar as they contribute to mental strength, but the question about the existence of degrees of consciousness is redirected towards the phenomenal intensity of states themselves. Naturally, when interpreted in this way, the question about whether there are degrees of consciousness is a resolute 'yes' because there certainly are degrees of mental strength.

It is unclear whether aggregating the mental strength of someone's set of conscious states at a given moment is what determines their degree of *global* consciousness. We get counterintuitive results if we accept this idea. Consider patients with unresponsive wakefulness syndrome (Laureys et al. 2010). These patients sustained severe brain damage and as a result they are in a state of partial arousal (e.g., they have circadian rhythms), but they seem otherwise unresponsive to commands and, therefore, unconscious. However, a subset of these patients are suspected to enjoy *some* conscious states despite their inability to respond behaviorally (Owen 2006). Imagine that one of their isolated conscious states had a very elevated mental strength, more elevated than all the states of a drowsy person taken together. On one hand, it may seems counterintuitive to say that an unresponsive wakeful patient is more conscious than a drowsy person who just woke up from a dreamless sleep. On the other, it is hard to think what else would be needed other than the mental strength of the aggregate of someone's individual states to determine their degree of global consciousness.

I suspect that part of the counterintuitiveness of this example stems from the fact that the drowsy person also has access-consciousness that the patient lacks. For instance, the drowsy person may be able to talk, attend, think, remember, respond, and perform all sorts of other tasks that require cognitive control. The patient cannot do most if not any of these

things. However, phenomenal- and access-consciousness tend to go hand-in-hand (Block 1995; Block 2007), which may be eliciting the wrong intuitions in cases like this one. Namely, perhaps our intuitions about the degrees of consciousness for global states mix phenomenal and access-consciousness, making it hard to accept an aggregation account of mental strength. Theoretically speaking, however, an aggregation account of mental strength might yield the right result that there are degrees of global consciousness and that they can be systematically ranked. In any case, I would point out that the notion of mental strength provides us with an extra tool to move fowrward this debate by refocusing the emphasis in the case of local consciousness and perhaps by reshaping the logical space of possibilities in the case of global consciousness.

d. The function of consciousness

Consciousness is often thought to perform important functions: flexible control of behavior, rational thought, and cognitive control (e.g., action inhibition and preparation, task switching, control of attention, working memory). Recently, philosophers (Phillips 2016) and psychologists (Peters and Lau 2015) have questioned whether perception is even possible without conscious awareness. Some philosophers have sometimes gone as far as to argue that consciousness is the mark of the mental (for example, Locke; see Coventry and Kriegel 2008).

There is, however, abundant evidence that some mental states, which often take place consciously, can also occur unconsciously. Subjects who fail to report awareness of stimuli can perform above chance in a wide array of visual and cognitive tasks, such as stimulus discrimination, word meaning extraction, simple arithmetic operations, and cognitive control in general (Dehaene 2014; Dehaene et al. 2014). Performance can be matched between more

conscious and less conscious conditions in visual tasks (Lau and Passingham 2006). Blindsight patients who have sustained damage to visual cortex areas can detect and discriminate stimuli they are unaware of (Weiskrantz 1986). In the most striking cases, blindsight patients can even avoid obstacles while walking down a hallway (de Gelder et al. 2008). Even when nuanced methods are used, researchers have often failed to demonstrate a clear advantage of consciousness (Koizumi, Maniscalco, and Lau 2015; Samaha et al. 2016). Moreover, many philosophers have forcefully argued that consciousness does not have a function, or at least that it was not evolutionarily selected for performing any function (Rosenthal 2008; Robinson, Maley, and Piccinini 2015). Together, this evidence suggests that many cognitive abilities exist (or could have existed) without consciousness, which puts pressure on the intuitive necessary link between consciousness and rational thought and action.

Despite these empirical findings and philosophical arguments to the contrary, it should be highlighted that mental strength performs at least two important functions: structuring the conscious mind and justifying self-guided action and reasoning. As argued above, mental strength structures the conscious mind. This structure has clear behavioral and cognitive effects. By prioritizing some states, mental strength guides action and cognition. Note that the fact that this could be done without consciousness and, therefore, without mental strength, does not mean that creatures like us can do it without consciousness or without mental strength. The same reasoning can be applied if consciousness and, thereby, mental strength, was not selected for (Robinson, Maley, and Piccinini 2015). That would not mean that in creatures like us mental strength does not perform an important function.

Conscious mental states do not have the same effects in our decision-making and in our mental lives independently of their strength. Certainly, there would be no reason to take an

aspirin if one were undergoing an unconscious pain (even if we were told by third-personal means that this was the case). There is less motivation for taking an aspirin when experiencing a mild headache than when experiencing a strong headache; a stronger emotion is easier to introspect than a mild one (see Chapter 2 for details on the relation of mental strength and introspection); a more vivid mental image is more useful for simulating a future scenario than a weak mental image; there is also less justification to take weak visual experiences at face value, and therefore act upon them, than when experiences are strong.

Philosophers and psychologists might have overemphasized the significance of the research discussed above, ignoring the role of mental strength in our lives: the initiation of action, the justification of perceptual beliefs, and the structuring of the stream of consciousness. Of course, an important motivational and justificatory role of mental strength is consistent with consciousness providing little to no advantage in visual processing (e.g., performance; see Chapter 4) as has been shown multiple times.

7. Conclusions

Mental strength is a distinct phenomenal magnitude of individual conscious mental states. It is what modulates the phenomenal intensity of conscious experiences, thereby modulating the degree to which mental states make their way to our consciousness. This important role of mental strength has been often underplayed by philosophers and psychologists in their theorizing and experimental designs. In this chapter, I offered an account of mental strength understood as a domain-general property of conscious experiences themselves, rather than as some aspect of their representational contents or attentional status. Instead of explaining

the degrees of consciousness in each domain by appealing to domain-specific representational and phenomenal characteristics, the theory advanced in this chapter offers a parsimonious account of the intensity of experiences by postulating the existence of a single domain-general intensity phenomenal property. Mental strength explains the synchronic and diachronic dynamics of the structure of the stream of consciousness. It also provides consciousness with a clear function in action and cognition. Finally, I showed that mental strength can be thought of in scientific terms and that we can reasonably speculate about its neural correlates.

Chapter 2

A Detection Theory of Introspection

In this chapter, I develop a theory of introspection in which mental strength plays a central role (see Chapter 1). By 'introspection' I understand the process of attentively focusing on one's current conscious mental events in order to form judgments about them. We can introspect perceptual experiences, pains, emotions, desires, and thoughts, among other mental events. Although there are many disagreements about introspection, I expect this definition to be relatively uncontroversial.[29] So defined, introspection should not be identified with consciousness, because not all conscious mental events are introspected. A theory of introspection also need not depend on any specific theory of consciousness, and this is true of the theory I will develop.[30] For concreteness, I will limit my examples to pains. The

---

[29] "[I]ntrospection is an attentive operation and one which is only occasionally performed, whereas consciousness is supposed to be a constant element of all mental processes […]" (Ryle 2009, 146); "When a thought occurs to you, or you make a conscious judgment, your attention is engaged.[…] A pain, for instance, can equally be an object of attention." (Peacocke 1998, 64-5); "[W]e can also direct our attention at our phenomenally conscious experiences as such, in introspection. I can pay attention to the way things seem to me while watching a game; or I can concentrate on the felt qualities of my experiences, noticing what they are like." (Carruthers 2000, 211; his emphasis); "[I]ntrospection is defined as deliberate and immediate attention to certain aspects of phenomenal experience." (Hatfield 2005, 279); "Introspection […] involves consciously and deliberately paying attention to our contemporaneous mental states." (Rosenthal 2005, 28); "The "organ" of introspection is attention, the orientation of which puts a subject in an appropriate relation to a targeted state." (Goldman 2006, 244); "I attend to my visual experience and think I am having an experience of such-and-such quality […]." (Chalmers 2010, 254) "Introspection is the dedication of central cognitive resources, or attention, to the task of arriving at a judgment about one's current, or very recently past, conscious experience […]." (Schwitzgebel 2012, 42); "Plausibly, introspection depends on a form of attention that enables selective thought about mental properties." (Wu 2014b, 254); "[A]n introspective state [is] structurally identical to a regular, non-introspective conscious state, differing only in respect of the distribution of a certain resource, which we may call attention." (Giustina and Kriegel 2017)

[30] My view is compatible with first-order (Block 2007; Lamme 2010), same-order (Kriegel 2009), higher-order (Rosenthal 2005; Lau and Rosenthal 2011), global workspace (Baars 1988; Dehaene and Naccache 2001) or information integrated theories (Tononi et al. 2016; Tononi 2008). In Chapter 4, however, I argue that the higher-order view is well supported by neural and computational data.

introspective model that I develop here depends on states having mental strength. The model, then, can be easily generalized to introspection of states other than pain (see discussion about the domain-generality of mental strength in Chapter 1).

It might help to start with some examples of when introspection is easy and when it is hard:

Accuracy. Unanesthetized dental patients can easily introspect their pains when their tooth is being drilled. However, patients undergoing dental treatment occasionally report feeling pain before the dentist's instruments even touch them. This occurs even in patients whose nerves have been removed or who had been anesthetized.[31] Surprisingly, patients do not report feeling pain once they are told that they had not been touched. Yet, they still insist it was painful the first time.

Detection versus discrimination. Patients can easily detect and report feeling a sudden pain. However, it is typically harder for them to discriminate between kinds of pain, such as whether a pain is quivering or shooting, gnawing or stabbing, located exactly at the center of the back or slightly skewed to one side.

Confidence. Patients admitted to the ER while experiencing a strong pain immediately and without hesitation inform the personnel how they feel. This same confidence, however, is not always displayed minutes after taking a potent painkiller. When asked if they are still in

[31] This case of dental fear is discussed by Rosenthal (2005, 127).

pain, patients may hesitate and when they finally answer, it is typically with lower confidence.

Criterion effects. When visiting the doctor, patients may detect mild pains that had been unnoticed when they were at home. Their criterion for classifying an experience as painful seems less strict in the presence of a doctor, and stricter when at home, perhaps because they want to make sure the doctor's diagnosis takes all of their pains into account.

In the next section, I will explain why existing theories of introspection cannot explain all these features of our introspective judgments. Then, I will introduce, motivate, and defend a new theory modeled after a widespread scientific theory of perception, Signal Detection Theory. Accordingly, I title my view Introspective Signal Detection Theory. I also discuss important differences between perception and introspection, as well as potential neural implementations of iSDT and ways of testing it.

The claim that introspection is similar to another cognitive capacity may be surprising. It used to be a common view "that the kind of knowledge a person has of their own mental (psychological) states, such as thoughts and feelings, is in principle not only fundamentally different from but also superior to the knowledge of their thoughts and feelings that is available to anyone else." (Alston 1971, 223) Access to one's own minds was thought to be peculiar (i.e., different) and privileged (i.e., superior or, according to some, infallible). I believe that this is a mistake and will argue that introspection is of the same general kind as the rest of our cognitive capacities.

1. Existing Theories

Most existing theories of introspection fall into two camps. Theories in the first camp try to explain why introspection can be infallible. Descartes vividly evokes introspective infallibility when he writes: "I am now seeing light, hearing a noise, feeling heat. But I am asleep, so all this is false. Yet I certainly *seem* to see, to hear, and to be warmed. This cannot be false" (Descartes 1985a, AT VII 29).[32] In this vein, some contemporary philosophers argue introspection of conscious experiences, such as perceptions, imaginations, pains, and thoughts, to be infallible—at least in some limited, paradigmatic cases (Chalmers 2003; Gertler 2001; Burge 1996). Some of these philosophers take introspection to be direct and self-verifying. According to Burge, if one judges 'I am thinking that my head hurts', "or indeed just engages in the thought, one makes it true. The thought is *contextually self-verifying*. One cannot err if one does not think it, and if one does think it one cannot err. In this sense, such thinkings are *infallible*." (Burge 1996, 92) According to Gertler, introspection takes place via pure demonstrative reference achieved via directing attention to the phenomenal contents of our conscious experiences: it is *thus* [here, now]. "By appropriately attending to the dull throbbing sensation [of a headache], you demonstratively pick out the phenomenal content <dull throbbing>." (Gertler 2001, 321) Thus, according to Gertler, phenomenal content is embedded in the introspective judgment 'it is *thus* here and now', preventing any sort of error. According to her, "pure demonstrative reference allows the subject to grasp the content *directly* […] in the sense that there is no causal gap between the referring state and its

[32] Whether Descartes indeed held a theory of the infallibility of introspection is contentious (Newman 2016).

referent, the phenomenal content. For the referring state instantiates the phenomenal content, by virtue of embedding its token."[33] (Gertler 2001, 323)

Theories in this camp might explain how introspection works in a small number of cases, specifically those in which the alleged target state is itself created or in which it is mentally pointed out. But they do not explain most cases. The embedding approach leaves unexplained a whole range of prototypical introspective judgments in which the contents are explicitly described, rather than just demonstrated. For example, it does not apply to judgments like "I'm hungry" and "The pain in my left leg is getting worse." Our introspective capacities are wider than simple "it is *thus* here and now" judgments.

Theories in the second camp argue that introspection is mostly unreliable and try to explain why. Schwitzgebel, for instance, argues that "we're prone to gross error, even in favorable circumstances of extended reflection" (2008, 259) about ongoing conscious experiences, including mental imagery, dreams, all kinds of perceptual experiences, pains, and cognitive phenomenology. For Schwitzgebel, a crucial indication of introspection's unreliability is that there is large individual and group variation in our introspective judgments, as well as widespread uncertainty while we make them. According to him, this shows that "we are both ignorant and prone to error" and that "we make gross, enduring mistakes about even the most basic features of our currently ongoing conscious experience (or 'phenomenology')." (2008, 247)

Schwitzgebel argues introspective judgments are the product of regular cognitive capacities such as perception, memory, attention, and cognitive control, among others. Thus,

[33] Note that Gertler (2012; 2018) defends a less ambitious view.

judgments of our experiences are "influenced by at least: expectations about my experience, my knowledge of the outward environment, my knowledge of what I can and cannot discern, culturally available metaphors and general theories about visual experience, and my knowledge of other aspects of my psychology." (2012, 33) However, it is unclear why other cognitive capacities, such as memory and perception, are often reliable, while introspection is always unreliable, even though introspection is supposed to be a product of them. Schwitzgebel does not provide an introspective-specific explanation of why introspection fails. I agree with him that introspection is not special, in the sense that introspection is not of a different kind from other cognitive capacities. However, we need specific explanations for why each capacity fails. For example, memory is not peculiar in any strong sense, but the explanation of why we fail to remember is different from the explanation of why we fail to see or hear.

There are theories that fall between these two camps. They usually identify specific conditions in which introspection is unreliable (Rosenthal 2005; Armstrong 1968; Goldman 2006; Reuter 2011; Giustina and Kriegel 2017). For example, Hohwy (2011) thinks we are unreliable when we introspect visual phenomenology. He argues that the neural mechanisms responsible for visual experiences become disengaged when the neural mechanisms for introspection are active, thereby weakening the experiences and making introspection less accurate. Bayne and Spener (2010) think that introspection is inaccurate when we include theoretical concepts in our introspective judgments.

However, these moderate theories do not provide a general account of introspection.[34] They limit themselves to providing psychological and neuroscientific details that explain

[34] For a notable exception, see (Goldman 2006, Ch. 9).

inaccuracies in *some* domain (e.g., visual experiences) or in some narrow conditions (e.g., when introspective judgments include theoretical concepts). It is unclear whether Hohwy's explanation for introspective inaccuracy of visual experiences can be extended to pains or emotions. Bayne and Spener's approach leaves unexplained cases of inaccurate introspective judgments that contain no theoretical concepts.

A satisfactory theory of introspection must identify and explain more than the conditions in which it fails. It must also identify and explain the conditions in which it is reliable (Spener 2015; Goldman 2004). For example, any theory of introspection must explain why we are reliable at detecting strong pains, not just why we are unreliable at detecting weak pains. It may be tempting to compromise and say introspection is fallible in hard cases and infallible in easy cases—such as introspecting 'perceptual simples' such as color and shapes or the presence of a strong pain (Allen-Hermanson 2015). However, a general theory of introspection ought to explain *with the same resources* why introspection is unlikely to fail in easy cases and likely to fail in hard cases. Rather than making easy cases exceptional, a satisfactory theory of introspection must offer a unifying explanation of its whole accuracy range. Current theories also do not explain the features of introspective judgments listed in the introduction. For example, they cannot explain why a dental patient may incorrectly report pain, or why it is harder for patients to discriminate the type of their pain than it is to detect the pain**.** They also do not explain in a systematic and general way why we are more confident when introspecting strong than mild pains, or why introspective judgments of identical pains may vary depending on criterion effects.

My goal in this chapter is to offer a new and more satisfying theory. According to Signal Detection Theory (SDT), the strength of a perceptual stimulus modulates perceptual accuracy. For instance, you are more likely to accurately perceive a person in an alley when

the alley is well-lit than when it is not. Here I defend the view that all conscious experiences have degrees of strength. A central tenet of Introspective Signal Detection Theory (iSDT), the theory that I advance here, is that the strength of our conscious experiences modulates introspective accuracy. For example, you are more likely to accurately introspect a strong pain than a mild pain. iSDT also accounts for confidence and criterion effects as well as for the difference between detection and discrimination during introspection. Since all conscious experiences have a degree of strength, as I argued in Chapter 1, the iSDT model is apt for explaining introspection in several domains beyond pain, such as perception, mental imagery, emotions, thoughts, and desires. Modeling introspection after a well-established method for measuring sensory sensitivity such as SDT promotes its study in a systematic way. Finally, abandoning introspection's alleged peculiarity and privileged status in favor of a naturalistic understanding has the potential of illuminating the underlying neural processes of introspection.

The comparison of introspection to perception may trigger some alarms, as philosophers often vilify this kind of comparison (Shoemaker 1996). Although iSDT may have *some* similarities to some so-called inner-sense theories of introspection, the details of my view are—as far as I can tell—novel and not reducible to a perceptual model (see Picciuto and Carruthers 2014 for discussion). Introspection shares with perception the fact that they create representations about their targets that are used to guide behavior. If I am right, introspection shares with perception the fact that the accuracy of our judgments depends on the strength of the relevant signal, whether it is an external signal (as in the case of perception) or an internal signal (as in the case of introspection). But these similarities do not imply that we *perceive* our conscious events during introspection and it should be noted that there are several important differences too. I discuss these at length in section 4.

In the next two sections, first, I offer a short primer on signal detection theory and, then, introduce iSDT.

2. Signal Detection Theory Primer

Here I introduce basic standard tenets of Signal Detection Theory. The reader well-versed in the topic may comfortably skip this section.

Imagine you walk by an alley late at night. If the alley's lamp is on, it would be easy for you to notice a man lurking behind the dumpster. His face will look bright and the contours of his facial features well-defined. If the lamp is off, his face will look dark and the contours of his face ill-defined. It would be hard for you to see this man and easy to take him for being just a shadow. Now imagine that, earlier, you read that a robber was on the run in your neighborhood. In this case, a quick glimpse to the dark alley might help you see the man's face even if it looks dark and grainy. With the robber on the run, a face-like shadow may be enough for detecting the lurking criminal.

The lesson from this simple example is that perception (detection and discrimination alike) depends on two factors, the signal-to-noise ratio of the perceptual signal created by the stimulus and on a detection criterion. When the alley's lamp is on, the well-lit stimulus creates a strong signal in your perceptual system. The noise from the stimulus (e.g., shadows on the man's face) is minimal. The signal-to-noise ratio is high and perceptual uncertainty is low. When the lamp is off, the dim face produces a weak perceptual signal and perceptual noise increases (e.g., there are more shadows in the man's face). In this case, the signal-to-noise ratio is low, and uncertainty increases. Importantly, perceptual detection involves

making a decision. This entails that, even with the same signal-to-noise ratio, you may still detect the dim face if your perceptual detection criterion shifts. When you learn that there is a criminal on the run, you need less evidence to detect the presence of a man in the alley. Your criterion or threshold for detecting a face is lowered when the probability of a man lurking in the shadows increases.

Let us now introduce some technical details. According to SDT, subjects detect (or discriminate) stimuli by comparing an internal response against a criterion placed along a decision axis (Macmillan and Creelman 2005; Green and Swets 1966) (Figure 1A). Stimulus presentation gives rise to an internal response in the subject's mind (the signal).[35] In easy cases, the stimulus may be large, bright, and on view long enough (Figure 1B, top panel). In hard cases, the stimulus may be small, dim, and viewable only for a very short moment (Figure 1B, lower panel). On average, easy cases give rise to a stronger internal response and hard cases give rise to a weaker response. In fact, this is what makes them easy and hard, respectively.

[35] SDT was originally developed as a mathematical tool to assess the accuracy of radars. It was only later that it was adapted to be used in psychophysics (Luce 1963; Green and Swets 1966). When used to assess perceptual sensitivity, the internal response is understood as a hidden psychological variable. SDT can be further adapted to describe the neural underpinnings of internal responses (Shadlen and Kiani 2013).

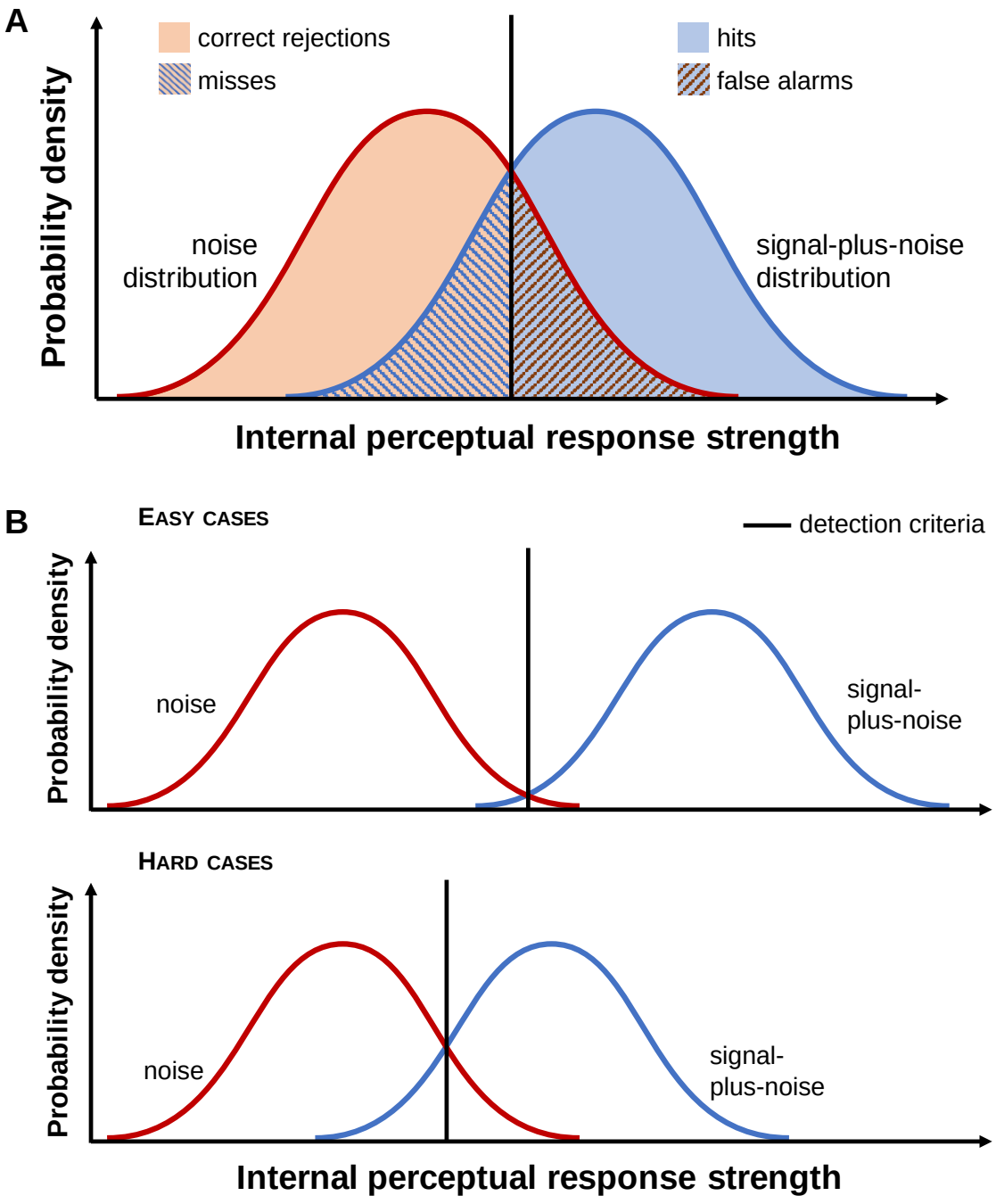

Figure 1. Signal detection theory

(A) Basic model of a detection task in Signal Detection Theory. (B) Top panel. Model of a perceptual task with easy trials. Lower panel. Model of a perceptual task with hard trials.

However, internal responses vary across trials and subjects for reasons other than stimulus properties (e.g., different sizes, brightness, viewable time, and stimulus noise). Variation in stimulus processing (e.g., attention, experience, natural skill, flukes, state of the perceptual system) and internal noise (i.e., internal response unrelated to stimulus presentation) also affect the signal-to-noise ratio of the internal response. Even when there is no stimulus, internal noise alone produces an internal response. When aggregated, the internal responses when the stimulus is absent (noise) and the internal responses when the stimulus is present (signal-plus-noise) form two normal (Gaussian) distributions. By setting a criterion along a decision axis, subjects compare the criterion to the strength of the internal response on a given situation and classify it as belonging to one or the other distribution. When the strength of the internal response does not cross the criterion, subjects respond as if the stimulus were absent. If it crosses the criterion, they respond as if the stimulus were present (Figure 1A).

Because there is always certain degree of overlap of the two distributions, there is always a certain degree of uncertainty in perception.[36] The smaller the overlap, the more different the noise-only and the signal-plus-noise internal distributions are and, hence, the more likely one is to respond correctly (i.e., making a hit or a correct rejection). When the distributions overlap more, detection becomes harder because a wider range of internal response strengths can equally belong to either distribution. So, in this case, even a criterion optimally placed to minimize errors generates a significant number of mistakes (i.e., false alarms or misses). The larger the overlap of the two distributions, the closer their means are and, conversely, the

[36] This uncertainty need not be reflected in the subject's subjective confidence. Subjects may feel completely confident about their perceptual decisions and yet the internal response evidence be ambiguous between pertaining to the signal or to the noise distributions.

smaller the overlap, the further apart they are. Thus, the distance between the means of the distributions can be used as a measure of detection sensitivity (called d').[37]

To put these technical notions into simpler terms: perceiving is never pure signal processing. Noise always gets entangled with the perceptual signal and perceiving consists, to a large extent, in recovering signals from ever-present noise. Perceptual sensitivity is the capacity of a perceiver to distinguish signal from noise. This is what it means to say that perceptual accuracy depends on the signal-to-noise ratio. The larger this ratio is, the more sensitive, i.e., the more accurate, a perceiver will be.

SDT also emphasizes that perception is a decision-making process. Sensitivity and response bias (i.e., criterion placement) are independent. Keeping the subject's internal response signal-to-noise ratio fixed, a detection criterion can be placed anywhere along the decision axis. Reward schedule, risk preferences, perceptual biases, or perceivers' priors can affect criteria placement (Witt et al. 2015; Macmillan and Creelman 2005). A subject with a liberal criterion (placed leftwards on the decision axis) would require very little internal response strength to get the subject to respond as if the stimulus were present. In contrast, a subject with a conservative criterion (placed rightwards on the decision axis) would respond as if the stimulus were present only when the internal response is rather strong. This explains why knowing that there is a robber nearby makes you more likely to identify the shadow in the alley as a person.

[37] The distance between the means of the distribution is sufficient for determining sensitivity assuming equal variance in both distributions. If these differ, the variance has to be taken into account.

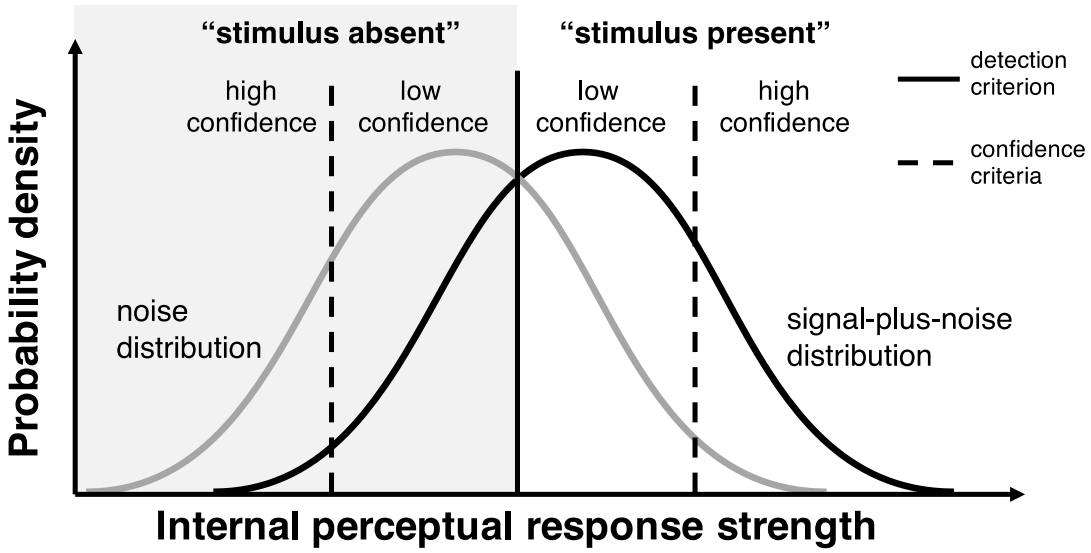

Figure 2. Confidence criteria in SDT

SDT also accounts for confidence in perceptual tasks (Figure 2). Confidence levels are determined by further criteria placed to the left and right of the detection criterion (Kepecs and Mainen 2012; Kepecs 2013; Shadlen and Kiani 2013; Macmillan and Creelman 2005). If the signal crosses the detection criterion ("stimulus present" in Figure 2) but not the confidence criterion, subjects report they detect the target with low confidence. If it also crosses the confidence criterion, they report detecting the target with high confidence. An analogous explanation applies when subjects do not report detecting a target ("stimulus absent" in Figure 2).

Finally, I note that SDT is primarily a theory of perception. SDT, however, can be adapted to explain the difference between conscious and unconscious perceptual processing (see Chapter 3). However, nothing in this section or about iSDT depends on this.

3. Introspective Signal Detection Theory

Introspective accuracy can be modeled after how SDT models perceptual accuracy. According to iSDT, introspective accuracy depends on an introspective signal-to-noise ratio. When one sets to judge the presence or the features of an ongoing conscious experience, the ensuing introspective judgment will be more or less accurate depending on the strength of the introspective internal response. This response depends on the degree of mental strength of the target conscious experience in the same way the perceptual internal response depends on the strength of the target external stimulus. There is also introspective noise, i.e., an internal response unrelated to the introspective response of the target conscious experience. Introspective noise can be introduced by fluctuations in attention and the mental strength of other experiences. A stronger experience generates a stronger introspective signal, which makes it easier to disentangle from noise and, consequently, more likely to be detected or discriminated accurately during introspection. In this section, I will develop the iSDT model focusing on pains. Because of the domain-generality of mental strength (Chapter 1), the model can be easily expanded to other experiences.

Let us recall the following pain example from Chapter 1. Imagine you wake up late for work. You stub your big toe against the bed when rushing to the shower. A sudden painful sensation invades your conscious awareness: your toe hurts. First, the pain is sharp, strong, and unpleasant. You hold your toe and while doing so you are unable to focus on anything else except the painful sensation. After a few seconds, your experience starts changing: it slowly becomes weaker while still being a sharp, unpleasant pain in your toe. As the pain decreases, your mind gets back to thinking about being late for work and you resume your plan to take a shower. As I argued before, an obvious phenomenal change takes place as the

intensity of your pain increases and decreases. This phenomenal change is due to a variation in the degree of mental strength of the experience.

Barring extraordinary cases, it would be surprising if one perceived inaccurately a large object that was looked at attentively for enough time under good viewing conditions. Similarly, strong pains are most likely introspected confidently and accurately (i.e., they are detected when present and not detected when absent, and their features are accurately discriminated). Failing to introspect a strong pain should be a very rare event. In contrast, it would not be surprising if one missed or misperceived a very dim, rapidly-presented stimulus while distracted or during other suboptimal viewing conditions. By the same logic, we should expect to introspect mild pains inaccurately, at least sometimes, and be less confident about our introspective judgments. The iSDT model provides a plausible and systematic account of why a strong pain is hardly ever missed, why it seems odd to think we could fail to introspect it, and why we are usually quite confident during introspection of strong pains. It also predicts that none of this holds for mild pains. iSDT is consistent with a natural understanding of pains and our knowledge of them, at the same time that it provides a psychologically plausible model of how introspection works.

Note that introspective inaccuracies are not accessible through introspection. At least not immediately and not always. Subjectively, if introspective judgments vary from one moment to another or if it is difficult to come to a verdict, one may conclude that introspection is amiss. Unlike perception, however, it is hard to get rid of an introspective mistake by non-introspective means (for example, by someone pointing out the mistake).[38] It might never

[38] It is worth pointing out that this only applies to conscious experiences. Introspective errors of thoughts or desires might be easier to detect by others, since they tend to last longer and they can manifest in behavior more prominently.

seem to oneself that an introspective mistake is taking place. This, I suspect, explains some of the resistance to the idea that we can be introspectively wrong about pain. So, rather than relying on introspection or third-party corrections, introspective inaccuracies are a consequence of iSDT and, thus, we must accept their possibility based (mainly) on theoretical reasons.

Now, I will show how iSDT provides a satisfactory explanation of the four features of introspective judgments listed in the introduction: accuracy, detection versus discrimination, confidence, and criterion effects.

a. Accuracy

Nothing about iSDT prevents the possibility that we are terrible introspectors. But we do not need to accept this sort of global skepticism (Schwitzgebel 2011). We can start by assuming that introspection, as the rest of our cognitive capacities, is accurate to a certain degree. Following iSDT, this entails that the overlap of the introspective noise and introspective signal-plus-noise distributions is from low to moderate, especially for strong target conscious experiences. Getting one's tooth drilled often produces an intense pain, that is, it often is a conscious experience with high mental strength. Thus, making an inaccurate introspective judgment about strong pains produced by getting one's tooth drilled is expected to be rare. But not impossible, because there is always some overlap between the noise and signal distributions (Figure 3A). On rare occasions, someone might make a false alarm judging they are in pain when they are not.

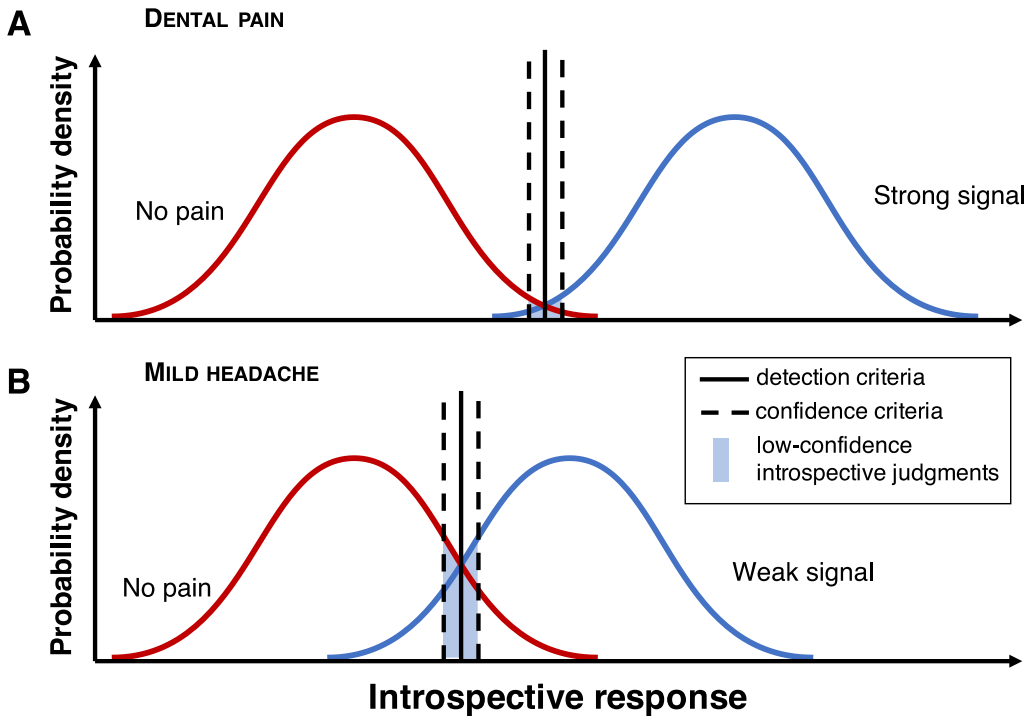

Figure 3. Introspective signal detection theory
(A) Introspective model of a stereotypical strong pain. (B) Introspective model of a stereotypical mild pain.

Recall the example of dental fear mentioned in the introduction. Vibrations produced by the dentist's instruments in conjunction with the patient's fear of the treatment produce a pain report even though anesthesia or the lack of nerves should make that impossible. In cases like these, there are at least two possible accounts of the patients' initial report. Patients in fact experience pain (and accurately introspect and report it) even if there is no obvious cause for it. Alternatively, patients do not experience pain and their reports are based on an inaccurate introspective judgment. For those still convinced pain introspection is infallible or that feelings of pain and introspective judgments of pain collapse into each other, the first answer would seem correct. However, the second explanation is simpler: subjects

misjudged their experience. Patients are not in pain (yet), but fear and the expectation of it propitiates constant introspecting their ongoing experiences. This constant evaluation eventually produces a false alarm. Maybe their criterion was more liberal or maybe by constantly attending their stream of consciousness they amplified the mental strength of a non-painful state, like fear, and eventually misclassified it as a pain.

iSDT also accounts for possible misses. The mental strength of a strong pain yields an introspective signal that most of the time crosses the introspective detection criterion. However, in some rare cases, the mental strength of an actual dental pain might not be sufficiently strong. Or perhaps the criterion is shifted to become more conservative than usual. This situation is expected to be rare for strong pains, just as failing to see a big object under optimal viewing conditions is rare. However, unlike missed big objects, introspectively missing a strong pain often remains uncorrected. This in part explains why it seems so counterintuitive that one could be in a strong pain and yet fail to judge that this is the case. Moreover, due to the professed incorrigibility of introspection (Dennett 2002; Rorty 1970; Langland-Hassan 2017; Alston 1971), any change in our introspective judgments is typically attributed to a change in the target conscious state rather than to a previous introspective error.[39] iSDT suggests this need not be the case.

Mild pains are different. They have less mental strength, which entails the introspective signal they produce is weaker. The noise and signal distributions of a mild headache overlap more, which entails we are introspectively less sensitive to the presence of mild pains and their features. In other words, on average, the mental strength of mild headaches is closer to

[39] This does not entail that phenomenal variability cannot be an acceptable explanation in some cases (Hohwy 2011).

the mental strength of introspective noise, making harder for us to detect and discriminate them. This does not necessarily entail that false alarms and misses are frequent, just that they are less rare than during introspection of strong pains (Figure 3B). Thus, we should expect introspection to be more accurate when the target is strong than when it is not. As can be easily noticed in Figure 3, the area under the red curve that crosses the criterion is larger during mild pains than during strong pains, indicative of more false alarms.

b. Detection versus discrimination

From SDT we know detection is easier than discrimination, which is easier than classification. This can be shown formally (Macmillan and Creelman 2005), but an intuitive example should suffice. In the dark alley, even a quick glimpse may reveal the presence of someone. Yet, a quick glimpse would not reveal whether the person in the alley is the robber or your neighbor. This point applies to introspection as well. Unless mental strength is extremely low, detecting the presence of certain mental state need not be very hard. Thus, introspecting accurately that one is in pain may be common, even for mild pains. But, as in perception, when mental strength is low we should expect more errors discriminating fine-grained properties of a pain (e.g., dull, throbbing, located exactly here or there, etc.). When experiencing a very mild pain, perhaps you are sure it hurts, but are you sure it hurts exactly *here*? Are you sure it is a throbbing pain and not a dull pain? If iSDT follows SDT in this respect too, we have the tools to explain why we can expect worse discrimination of mild pains' features than those of strong pains.

Given this explanation, it is not surprising that philosophers find examples of introspecting being cold or seeing a red tomato as strong evidence in favor of infallibility,

while they use unconstrained cases of introspecting a mental image of one's childhood house for arguing in favor of its unreliability. In the former cases, the experiences are typically strong and they ask to detect a state or discriminate a simple property; in the latter, the experiences are unstable, complex, often weak and the examples require us to identify complex features of the image.

c. Confidence

When you stub your toe against the bed, at the beginning at least, the pain dominates your stream of consciousness. If you were to entertain a thought about it, the presence and nature of such a conspicuous and dominating event should be readily known and you should be confident about your judgment. When the pain in your toe is about to vanish, introspection is not as easy, and you are not as confident when judging it. You may be unsure if what you are feeling is still pain or not. You might be aware of a sensation, but is it still a pain? Is it sharp, throbbing, or dull? Where exactly does it hurt? Is it the toe, just part of the toe, or part of the instep too? You may be able to tell it is not very painful anymore and, perhaps, you may identify it as dull and as located in the tip of the toe. Perhaps you are even right, but you are unlikely to be as confident as before. Note that this is how one can accurately determine that a state is weak. The fact that one is not able to introspect it easily and confidently reveals its weak nature. The situation is not different from perception. If a dim stimulus is flashed very quickly on a screen, you can very accurately determine that it was a weak stimulus, even if you fail to identify any of its traits (or precisely because of it).

The existence of this difference has been recently confirmed by a linguistic analysis of internet searches according to which, English and German speaking users use "I feel pain"

more often when describing minor or little pains, while they use "I have pain" significantly more often to describe severe or major pains (Reuter 2011). Presumably, subjects follow the usage of feel/have in a similar way as in other modalities to express degrees of confidence indicative of an understanding of an appearance/reality distinction. Compare: "the shirt looks blue" vs "the shirt is blue," where the former is used to express low confidence about the real color of the shirt and the latter expresses confidence in the perceptual judgment. Similarly, subjects use "feel pain" for mild pains to express their weak confidence in their reported experience, while they use "have pain" to express certainty about the presence and characteristics of their experience.

iSDT provides a systematic explanation of this confidence variability: stereotypically strong pains, such as getting one's tooth drilled, have fewer instances of introspective internal response falling between the detection and the confidence criteria than weak pains, such as mild headaches (Figure 3; shaded region).[40, 41]

---

[40] Snodgrass and Shevrin (2006) describe a similar situation; their focus, however, is on perceptual internal response, not on what I call here introspective response. According to them, trials whose internal response fall between a detection criterion and a subjective criterion (akin to a confidence criterion [see Chapter 3; Figure 6]) can be labeled as "weak consciousness trials." According to them, these are phenomenally conscious trials that are, however, not access-conscious trials (Block 2007; Block 1995). On their view, this would entail that they are not amenable to introspection, which requires the orientation of attention and forming a judgment (but see the end of section 4 below). My view is different from theirs because I think that trials that give rise to experiences with little mental strength, and therefore little introspective response, can still be introspected, albeit more inaccurately so. For a criticism of Snodgrass and Shevrin's view, see (Irvine 2009).

[41] The extent to which mental strength and introspective confidence influence metacognition and confidence ratings in perceptual tasks is of great interest, but I do not discuss it further here. See (Morrison 2016; Morrison 2017; Denison 2016; Fleming and Lau 2014; Overgaard and Sandberg 2012).

d. Criterion effects

Criterion effects can account for introspective variation too, even when holding introspective sensitivity fixed. The example of patients who detect a pain depending on whether they are in front of a doctor or not can be explained by shifts in their introspective criteria. By shifting what degree of mental strength counts as pain, these patients fail to detect it when they are at home. In contrast, the patient with dental fear illustrates a case of someone with a liberal introspective criterion. Being scared makes the patient count as pain even a weak introspective response pertaining to noise. Both cases propitiate introspective inaccuracies.

Applying the assumptions of SDT to introspection is a reasonable and fruitful endeavor. iSDT provides a robustly grounded framework for thinking about our introspective capacities and mental strength is the kind of signal these can track. Important features of introspective judgments that make these hard are fully accounted by iSDT.

Despite the different types of phenomenal and representational components of experiences like pains, perception, or mental imagery, on one hand, and of thoughts and desires, on the other hand, it is not unreasonable to say that *qua* conscious states they all have a degree of mental strength (see Chapter 1). Due to the centrality of mental strength in iSDT, the domain-generality of mental strength allows us to apply the same lessons learned about introspection of pains to other mental states. Naturally, details about introspection of other types of state need to be fleshed out, but such endeavor will need to be sought in another occasion

4. Perception and Introspection

SDT provides a helpful model for understanding introspection. Yet, the parallel between perception and introspection, and consequently between SDT and iSDT, is not perfect. For example, the assumptions behind SDT, such as the Gaussian shape of the internal response distributions or the independence of sensitivity and response bias, have been empirically validated by more than half a century of psychophysics (Luce 1963; Swets 1961; Green and Swets 1966; Macmillan and Creelman 2005). They have also been validated by a wide variety of experiments involving visual, auditory, and haptic perception and even memory (Wixted 2004). This speaks to the generality of these assumptions. However, it remains to be determined whether they transfer to mental strength and introspective judgments as proposed by iSDT.

Setting these technical details aside, many philosophers resist comparisons between perception and introspection (Shoemaker 1996). I agree with these philosophers that we do not *perceive* our mental states when we introspect them. In this section, I will list some of the most important differences between perception and introspection.

The first difference is about the appearance/reality distinction. In perception, there is a clear distinction between reality and appearances. You can be wrong about seeing a light or hearing a sound. However, some deny that you can be wrong about whether there *seems* to be a light or whether there *seems* to be a sound. Recall Descartes's claim: "I certainly *seem* to see, to hear, and to be warmed. This cannot be false." Conscious experiences *are* the appearances, and, it is argued, there cannot be a further appearance/reality distinction involving these appearances. If there were one, there would be appearances of appearances. But an appearance of an appearance is indistinguishable from the first appearance.

If this is true of visual and auditory experiences, it is certainly true about pains.[42] Some may think that it follows that there is no difference between the appearance of being in pain and being in pain. According to this line of reasoning, appearing to be in pain should collapse into being in pain. But then we could not be introspectively wrong about being in pain. Even when you are not originally in pain, if it introspectively seemed to you that you are in pain, your seeming to be in pain should collapse into actually being in pain. Contrary to one of iSDT's important tenets, this would make introspective failures impossible.

This might lead some to doubt that there is an appearance/reality distinction in introspection, and thus whether iSDT is tenable. But this rests on an ambiguity in the term 'seeming'. It is one thing to have an experience such that things "phenomenally seem" to you to be a certain way. It is another thing to make a judgment such that it "epistemically seems" to you that something is the case.[43] In both cases, things *seem* to you to be certain way, but they are very different in nature. Nothing about these two types of seeming necessitates a collapse of one into the other. Judgments about pains (epistemic seemings) do not collapse into painful experiences (phenomenal seemings). Thus, an appearance/reality distinction in introspection can be preserved.

This appearance/reality structure allows for introspective inaccuracies. According to iSDT, epistemic seemings (appearances) need not match your phenomenal seemings (reality). Importantly, subjects appear to be sensitive to this appearance/reality distinction when they evaluate their pains. As noted in section 4.*c*, when people consider that their introspective judgments may be departing from their actual phenomenal experiences, they use expressions

---

[42] It is a contentious issue whether pains can really be assimilated to perceptions. Hence, it is debatable what counts as the underlying reality of which pain experiences are appearances of (Aydede 2009).

[43] See (Schwitzgebel 2008) for discussion of this point.

like “feeling a pain.” This indicates they suspect a mismatch between their experience and their introspective judgment of the experience. In contrast, they use expressions like “having a pain” to indicate they are confident that there is no mismatch between their pain and their introspective judgment of the pain. In other words, they are confident that their epistemic-seeming expressed by an introspective judgment matches their experiential reality.

Although there is an appearance/reality distinction in introspection, it is worth highlighting that its structure differs from that of perception. Conscious perception has a three-item structure: (1) the external object (i.e., the external reality), (2) the experience (i.e., the appearance or phenomenal-seeming) and (3) the perceptual judgment (i.e., the epistemic-seeming). Introspection has a 2-item structure: (1) the experience (i.e., the internal reality or phenomenal-seeming) and (2) the introspective judgment (i.e., the appearance or epistemic-seeming). These differences do not affect the main tenets of iSDT.

The second difference has to do with phenomenal character. In normal circumstances, conscious perception gives rise to a phenomenal character related to its object. For example, there is a special phenomenal character that accompanies perceiving an object as red. In contrast, introspecting a conscious experience does not have a phenomenal character related to the target conscious experience. An introspective judgment of pain does not have a phenomenal character related to the pain or to painfulness. If it did, inaccurate introspection would give rise to incompatible phenomenal characters (e.g., simultaneous phenomenal characters of having a sharp and a dull pain in the same bodily location). But, to my knowledge, this kind of conflicting phenomenology does not take place.[44]

---

[44] Experiencing “impossible colors” is the closest case I can think of. Due to physiological constraints of our visual system, we do not perceive reddish-greens or yellowish-blues. In contrast, it is possible to see blueish greens or yellowish reds. However, under certain experimental conditions, some subjects

It is possible that introspective judgments have a *distinct* phenomenal character, as defenders of cognitive phenomenology would argue (Montague 2015). In stark contrast to perception, however, the new phenomenal character would be related to the act of judging, not to its target. Thus, introspecting a pain may have a particular phenomenal character provided it is not the phenomenal character of being in pain (for the reasons offered in the previous paragraph). Note that this is true too of perceptual judgments. In the Müller-Lyer illusion, you experience—it phenomenally seems—that one line is longer than the other. But you do not perceptually judge—it does not epistemically seem—that this is the case. This perceptual judgment, however, does not alter your *visual* phenomenology. You still experience the lines as having different lengths.[45] Likewise, when you judge that you are in pain, this does not alter the phenomenology of the pain itself, even if there is a phenomenology of judging.

Some may argue there is a third important difference: the outputs of perception and introspection are different. Perception produces representations below the level of judgment, and introspection produces judgments that require the possession of concepts. In that case, if we use SDT to model perception and iSDT to model introspection, the theories may not be sufficiently parallel because they have different kinds of output. I do not dispute that perception and introspection may have different outputs. However, this difference is not relevant for the points advanced by iSDT. In particular, the role played by signal strength in perception and introspection is informative regardless of their outputs. Moreover, it is

---

report seeing an area in their visual fields as simultaneously green and red (Crane and Piantanida 1983). Even if taken at face value, this result does not show that one can experience being in pain and not being in pain at the same time.

[45] Defenders of cognitive penetration of perception might disagree that this holds across the board (Wu 2017; but see Firestone and Scholl 2015).

possible to understand SDT as a theory of perceptual judgments. It is also possible to understand introspection as a capacity for entertaining epistemic seemings below the level of judgment and reformulate iSDT accordingly. Thus, the difference in outputs is not indicative of some irreconcilable difference between SDT and iSDT. I briefly discuss what these changes would look like.

On one hand, the machinery of SDT can be used to explain perceptual judgments, rather than perceptual representations. After all, SDT does not have the fineness of grain to distinguish between proper perceptual biases and response biases, and it is equally useful for both (for discussion, see Witt et al. 2015). That the formal apparatus of SDT can be adapted for iSDT, which is at the level of judgment, speaks to this flexibility. This does not mean that the outputs of perception are not perceptual representations.

On the other hand, it is possible that introspection takes place without the involvement of judgments. This would happen if there are introspective epistemic seemings below the level of judgment. Consequently, we could reformulate iSDT as a theory of introspective epistemic seemings rather than as a theory of introspective judgments, which is what I did here. An animal devoid of the concept of pain may still have an epistemic seeming of its pain phenomenology such that it is mistaken about it. Such animal might overestimate how much pain it is experiencing in ways that affect its decision-making without ever formulating an explicit introspective judgment or without the possession of concepts.[46]

[46] For discussion of possession of concepts by nonlinguistic animals see (Allen 1999).

Let me finish this section by iterating that the differences between perception and introspection discussed in this section do not affect iSDT. The similarities between perception and introspection that I drew upon to develop iSDT are related to the detection and discrimination of signals, not to their outputs, their phenomenal characters, or their appearance/reality structures.

5. The Transparency of Experiences

Philosophers often quarrel about whether the phenomenal character of experiences can be introspected (see Chapter 1 for discussion about the relation between mental strength and representational contents). According to some (Harman 1990; Tye 2000; Tye 1992; Shoemaker 2000; Dretske 1995), introspection does not reveal any intrinsic, non-representational, immediately 'felt' quality of experiences because these are transparent. This means that when introspecting a conscious experience, you really focus your attention on the worldly properties and objects the experiences are about, rather than on properties of the experiences themselves, as if these were transparent. According to this view, "no new features of your experience are revealed" during introspection, just "qualities of things in the world (as in the case of perceptual experiences) or of regions of our bodies (as in the case of bodily sensations)." (Tye 2016) Accordingly, when you introspect your toe pain, you really just focus on your toe rather than on some intrinsic, qualitative property of the painful experience (Tye 1995). Similarly, when you introspect the experience of seeing the deep blue of the ocean while standing on the beach, the focus of your attention is on the ocean itself and its worldly blue, not on some intrinsic, blue quality of your experience (Tye 1992).

There are voices of dissent among an apparent consensus around transparency (see Kind 2003). Philosophers sometimes appeal to raw feelings that are not supposed to be representational, like orgasms or moods, to counter transparency (Block 1996). If these sensations are not about something in the world, they cannot be transparent and introspection must focus directly on them rather than on outer objects. Sometimes the appeal is to cases of experiences that are qualitatively different without any representational difference. For example, the experience of seeing a square as a square or as a diamond (i.e., a tilted-square) or the experience of seeing a 3×3 grid of dots as three rows or as three columns (Peacocke 1992). The idea is that we could only know about these qualitative differences if introspection picks out intrinsic properties of experiences since representational contents (i.e., worldly objects and properties) are identical. Another preferred example is blurriness. Experiencing blurred vision of an object and an experience of sharp vision of that same object blurred could have the same representational contents, at least in some carefully matched cases, and yet having these experiences feel different (Boghossian and Velleman 1989).

Despite these efforts, defenders of transparency adamantly reject them, often finding a way to argue that the purported difference in qualitative character can be reduced to a difference in the representational contents of the experiences (Tye 2002; Tye 1992). This view, known as representationalism, can be summarized in the slogan "If two experiences are alike representationally, then they are alike phenomenally (and vice versa)."[47] So, the representationalist argues that the contents of the experiences of the two squares are different (and hence their phenomenal characters are different too). In one case, the content is something like <vertical square> and in the other is <tilted square>. By making the

[47] See note 21.

orientation part of the content, the difference in phenomenology is reduced to a difference in representational contents. And thus, the difference revealed by introspection is explained by differences in the world, not by directly accessible and intrinsic properties of the experiences. The other cases are similarly addressed.

Considering mental strength offers a novel challenge against transparency and against strong versions of representationalism. Mental strength not only modulates introspective accuracy, it is also available to introspection. In perception there is no tension saying that brightness improves the discriminability of other stimulus properties and that we can estimate the brightness of a stimulus. In introspection, there should be no tension either. Mental strength promotes accurate introspection at the same time that it is an aspect of experiences available to introspection. We can estimate the strength of experiences (for example, by ranking them). But mental strength is a phenomenal magnitude and as such it cannot be a property of external objects. It is intrinsic to experiences.

When attending to the strength of an experience it is not the world or its properties that we focus on. When describing his being transfixed by the intense blue of the Pacific Ocean, Tye writes: "I experienced blue as a property of the ocean not as a property of my experience. My experience itself certainly wasn't blue." (Tye 1992, 160) It is probably true that his experience was not blue. But it does not follow that his experience itself was not strong. This is probably what explains why he was transfixed by it. But the strength of his visual experiences is a property of Tye's mind, not a property to be found in the ocean. Shifts in his stream of consciousness could alter the mental strength of the visual experience of the ocean without altering its representational contents. Hence, with respect to their strength, experiences are not transparent. And if they are not transparent in part because the phenomenology of mental strength is not reducible to representational contents, then

representationalism is not generally true because it does not hold for all kinds of phenomenal characters.

6. A Science of Introspection

An important advantage of iSDT is that it puts introspection on the same standing as other cognitive mechanisms and, as such, it is amenable to scientific investigation. Here I outline how to move forward.

a. iSDT psychophysics

For some, studying introspection from the third-person point of view is a non-starter. While scientists precisely control stimulus properties in their study of perception, they cannot precisely control conscious experiences in their study of introspection. This seems to make subjective judgments the last word regarding someone's conscious experiences. Even if true, this does not entail that each subjective report is accurate or consistent with others; like perceptions, they can be more or less accurate. A comparison may help. The best way of knowing the length of an object is by measuring it. This does not entail that we can ever know the true length of an object or that every time we make a measurement we will obtain its true length. The consistency among measurements will largely depend on the reliability of the measuring device. By making several measurements under different conditions, a better estimate of the length of the object and of the reliability of the ruler can be obtained. Similarly, by measuring multiple times subjects' introspective judgments under different

conditions, we may estimate what their experiences are like and how reliable their introspection is, even if we do not have access to their true experiences. Thus, we can bootstrap our way above the authority of individual subjective reports. In this subsection, I argue that the assumptions from iSDT can be used to develop a measure of introspective reliability.

Some of the abilities that allow us to navigate the world (both spatially and epistemically) are more introspection-reliant than others and they can be exploited to study introspection (Spener 2015; Chirimuuta 2014). Presumably, introspection helps us succeed at many tasks. Examples of these include focusing binoculars, ordering a certain size of pizza depending on how hungry we feel or determining whether it is time to unbury our hand from the snow before it becomes too painful. Some of these capacities can be used in experimental conditions to test for variations in introspective sensitivity and validate some of iSDT's assumptions.

Tasks where subjects are asked to compare painful stimuli can be more reliant on introspection than others.[48] Imagine a task in which electric shocks are delivered via two electrodes attached to subjects' left and right legs, respectively. On each trial, the experimenter delivers a painful electric current. One leg, counterbalanced across trials, has a fixed, intermediate shock intensity. The other one starts off at a random intensity within a predefined tolerable range. Subjects' task consists in turning a knob that modulates the intensity of the variable current until they deem both of them equally painful. This SHOCK/SHOCK TASK is introspectively undemanding because subjects just need to judge whether the two pain intensities are the same or different overall without focusing their

[48] I adapt these examples from the visual tasks discussed in Chirimuuta's (2014).

attention on each pain independently. Now, imagine a similar task except that instead of the fixed-current electrode, one of the subjects' legs is attached to a patch that delivers a fixed heat-based painful stimulation. Like before, subjects modulate the electric current of the patch attached to their other leg until they judge the pains in both legs to be of equal intensity. In this HEAT/SHOCK TASK, introspection is more demanding because subjects have to discount the phenomenal differences of the two sources of pain. In Chirimuuta's words, it requires the "capacity to analyze and compare sensory experiences that bear nonobvious relationships of similarity and difference to each other" (2014, 917). Perhaps heat pain is more distributed and electric pain is more focalized. Maybe heat pain feels burning while electric pain feels stinging. Subjects, then, are required to focus on the independent dimensions of pain intensity and type, finding an identity along one dimension and disregarding the difference in the other.

We can take advantage of the introspectively demanding HEAT/SHOCK TASK to measure introspective variability. According to iSDT, if subjects' painful experiences have higher mental strength, their introspective judgments should be more accurate. Assuming the target states are similar, the judgments should also be less variable. Imagine an experiment with two conditions. First, a strong mental strength-inducing condition that uses stimulation well-above the detection threshold (say, the fixed heat patch is at 75% of the intensity range). Variations could include: sufficiently long lasting electric and heat stimulations, large enough contact areas, or fully attentive subjects. Subjects proceed to match the intensity of the electric and heat pains in tens, perhaps even hundreds, of trials. Since the heat patch delivers a fixed stimulation, if introspection is reliably accurate, one should expect subjects getting close to adjusting the knob to the same intensity on every trial (assuming habituation and tiredness effects are controlled for).

By iSDT's own assumptions, we can expect variations in internal introspective strength on every trial and, hence, variations in their knob settings. These variations may take place even if subjects are generally speaking introspectively reliable. That said, just like in perception, when the signal is strong enough, accurate subjects tend to be quite consistent.[49] Their small degree of variability, however, could be used as a baseline against which matching results from a weak condition are compared. A weak mental strength-inducing condition is achieved by lowering the heat-patch stimulation to the lower 25% of the stimulation range. Variations could include: introducing time pressure, reducing the contact surface areas, increasing subjects' cognitive load, or distracting them in some other way. In this weak condition, painful experiences should be harder to introspect according to iSDT. In consequence, more variability is expected across trials in the estimation of the current's intensity required for matching the heat and shock pain intensities. We then compare the mean and variance of subjects' estimations to the baseline mean and variance obtained in the strong condition. The difference is used as a proxy of the extent of introspective accuracy modulation exerted by mental strength both within and across subjects. Additionally, subject responses can also be accompanied by confidence ratings, providing a chance to further assess their metacognitive access to their answers (Fleming and Lau 2014).

Weak and strong trials could be interspersed to avoid habituation. Limiting stimulation to the 25%/75% points of the stimulation range prevents flooring and ceiling effects. This also avoids well-known issues detecting differences at low and high ends of the stimulation

---

[49] The technical reason for this is that when subjects are more sensitive, the noise and signal distributions are further apart. Even if the same stimulation gave rise to some variability in the internal response, in most trials this would not be enough to make it be on a different side of the criterion. In contrast, when subjects are not sensitive, the distributions overlap more and a small variation in the internal response may put it on a different side of the criterion changing the decision's outcome.

spectrum (e.g., it is easier to detect a 1g difference in a 10g object than in a 10kg object).[50] Besides, the difference between the low and high points should be just a few degrees/amperes, namely, small enough to avoid creating a sensitivity to large differences but sufficient for producing a difference in mental strength and consequently in introspection. Or at least this is iSDT's prediction. The proposed experiment is just a sketch and, even if all the necessarily details were fleshed out, it might not yield the results anticipated by iSDT, making it empirically inadequate. The important point, however, is precisely that iSDT makes empirically testable predictions.

b. iSDT and the brain

In this subsection, I explore a possible neural implementation of the introspective mechanisms proposed by iSDT. As above, this exercise is highly speculative and I do not strive for thoroughness or strict neurological accuracy. Rather, I will use a series of simple

---

[50] Philosophers often invoke Weber-Fechner's law, according to which increment thresholds (just noticeable differences or jnds) are proportional to stimulus magnitude approximated by a logarithmic transform. Philosophers seem to forget, however, that "Fechner's derivation of his law has been subjected to examination and criticism that are perhaps without parallel in the history of psychology." (Savage 1970, 290) Weber-Fechner's law does not hold empirically across all stimulus types and across all conditions. For instance, the law seems to work relatively well for weight in most conditions, but not for sounds, and it tends to break at higher levels of stimulation. Different sources of stimulation of the same type have difference power functions that account for the observed perceptual scale. Thermal pain increases more or less linearly with stimulus increases, while pain induced by electric shocks has an expansive power function, namely, pain units increase proportionally much more than stimulation increases. To put it simply, while a little bit more heat produces a little bit more pain, a little bit stronger electric shock produces a lot of more pain. Finally, appealing to internal noise is in fact one of Signal Detection Theory's fatal blow to the law. Perceptual scales derived from gathering just noticeable differences, which Weber-Fechner's law expects to be proportional to stimulus magnitude, are imprecise due to the accumulation of noise as jnds are integrated along the perceptual scale (Macmillan and Creelman 2005, 22-24; Kingdom and Prins 2010, 202-205; Savage 1970, 283-363; Luce 1990, 73).

models (not without plausibility) to show that iSDT makes claims that are amenable for neuroscientific testing and that its assumptions are within reasonable parameters. I anticipate the general conclusions I will draw. First, when presented with a stimulus, the brain generates an internal sensory response that, along with the representation of signal and noise distributions and the setting of a criterion, produces a sensory decision (e.g., detection, discrimination, classification, etc.). Second, when these sensory decisions are conscious, they have degrees of mental strength. Third, analogous to the sensory process, the mental strength of conscious experiences gives rise to an internal introspective response that, along with the representation of introspective signal and noise distributions and the setting of an introspective criterion, produces an introspective judgment (e.g., detection, discrimination, classification, etc.). In what follows I will, first, address how pain intensity judgments, as proposed by SDT, may take place in the brain. Then, I address how introspective judgments about the intensity of painful experiences, as proposed by iSDT, may be implemented.

An external stimulus, such as an electric shock, creates an internal sensory response in the brain (Figure 4, red sphere). The brain maintains a representation of the relevant noise and signal distributions to which that particular token response belongs to. Then, the sensory system evaluates whether the internal response crossed the detection criterion or not and it is classified as belonging to one of two distributions, either noise or stimulus present (Figure 4, purple sphere).

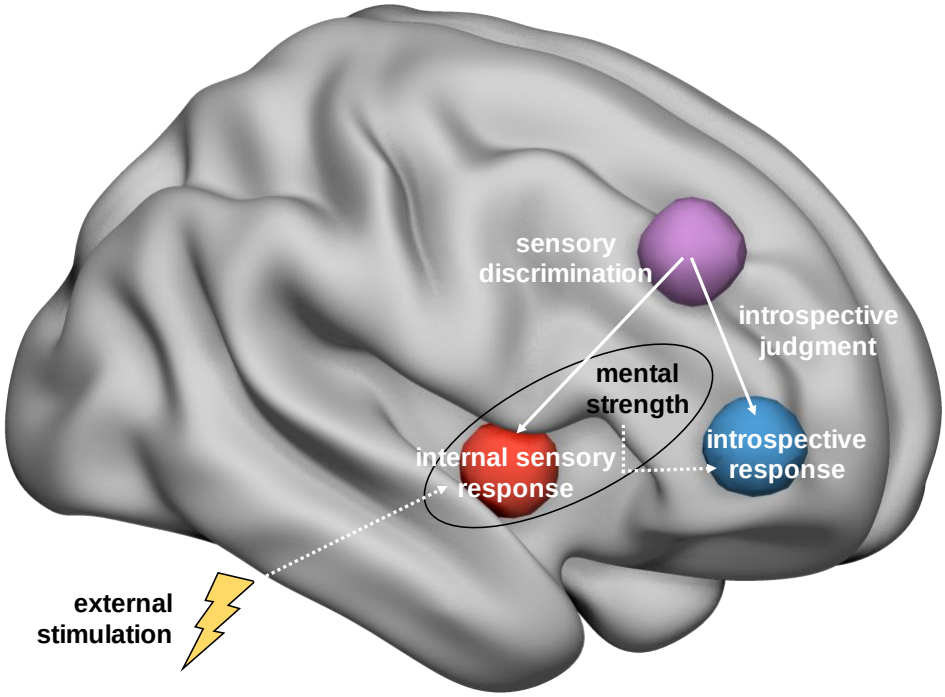

Figure 4. SDT and iSDT neural models

A few clarifications are in place. First, the internal sensory response need not be localized in a single area. As much research has shown, neural responses to pain intensity are distributed in widespread networks (Coghill et al. 1999; Atlas et al. 2014; Wager et al. 2013). This may or may not generalize to other sensory experiences. Second, the internal sensory response need not be itself the (total) neural correlate of conscious pain. Consciousness likely requires the interaction of pain intensity encoding areas with other regions, possibly in a frontoparietal network (see Chapter 4; Odegaard, Knight, and Lau 2017; Atlas et al. 2014). In consequence, pain mental strength—which by definition is conscious—may also be distributed in several areas, some of which do not code for pain strength itself, but consciousness in general (as represented by the shaded oval in Figure 4). Third, there are

different options of how the brain may represent the noise and signal distributions. One option is the "same-locus" view. According to it, the distributions are represented in the same areas that process the stimulus itself, that is, where the internal sensory response is generated (Figure 4; red sphere). By encoding in the sensory representation of a stimulus the probability distributions to which it belongs, the brain would effectively encode the internal response and its likelihood ratio in a single representation (Kiani and Shadlen 2009). This is compatible with the view that sensory experiences themselves assign degrees of confidence by including probability distributions as part of the experience itself (Morrison 2016). Alternatively, according to "different-locus" views, the brain may encode the trial-wise strength of the internal response in one area (say, the red sphere located in the insula; Figure 4) and the distributions for that particular kind of stimulus in some other area (say, the violet sphere located in prefrontal cortex; Figure 4) (Odegaard, Knight, and Lau 2017; Rahnev et al. 2011 are candidates for holding this view). The brain would make sensory detection and discrimination by situating the current experience's internal response in its corresponding position of the x-axis of the signal and noise distributions.

Sensory decision criteria (as well as confidence criteria) may be implemented by means of changing the excitability, the baseline firing rate, or the evidence accumulation rate of the relevant neural populations (Mulder et al. 2012; van Ravenzwaaij et al. 2012). Either way, the subject's sensory detection threshold would be affected, effectively changing the criterion. These neural changes could take place in the same areas encoding the internal sensory response and the signal and noise distributions. This would entail that most, or even all, of the relevant sensory processes take place in a single sensory area. Alternatively, in accordance with the different-locus view, the criterion may be implemented in the area where

the signal and noise distributions are encoded. Decision criteria may also be implemented in a third, independent decision area.

The neural implementation of introspection may follow an analogous structure. The regions responsible for encoding mental strength give rise to an internal introspective response (say, somewhere in prefrontal cortex; Figure 4, blue sphere). Like above, the introspective signal and noise distributions for that kind of experience may be encoded in the same or in a different region from where the introspective internal response tracking mental strength is kept. If a different region from where introspective internal response is encoded is involved, two possibilities emerge. One option is that this region coincides with the same region responsible for keeping track of the sensory internal response. This option is suggested, for illustration purposes, in Figure 4 (violet sphere). A second option would be that a new, independent region is involved. Introspective criteria are likely to be implemented in a similar fashion as sensory criteria. Subjects would make introspective judgments by placing the introspective internal response in its corresponding place within the introspective response axis. As they classify the experience as pertaining to the signal or noise distributions by comparing its strength to an introspective criterion, the conceptual machinery required for making judgments would be recruited. This whole process results in subjects making introspective judgments.

This simple sketch of a possible neural implementation of iSDT shows that introspection is subject to neuroscientific research. Together with the possibility of doing psychophysical testing for introspective variability, a clear picture emerges: the main tenets of iSDT are subject to empirical behavioral and neuroscientific research.

7. Conclusions

Mental strength, the phenomenal magnitude of conscious experiences, is a distinct aspect of their phenomenal character. Mental strength plays a crucial role in our mental lives: it modulates the accuracy of introspection. iSDT explains why sometimes we can expect to be introspectively very accurate, in line with infallibilists such as Burge or Gertler, and why sometimes we should expect to be inaccurate, in line with skeptic worries like the ones voiced by Schwitzgebel. In other words, it validates the intuitions of extreme, incompatible views. I take this to be a virtue of the theory. Unlike other theories of introspection, including moderate ones, iSDT offers a detailed, systematic, naturalistic, and psychologically plausible explanation of cases of various sorts. It also offers clear predictions of introspective performance under a wide range of conditions, and it postulates a general, testable psychological mechanism and plausible neural implementation of introspection. Instead of focusing on a few cases that may give rise to errors, as most moderate views do, iSDT fixes the whole range of success of introspection, including not just odd cases where it fails, but also common scenarios where it succeeds. Importantly, it achieves this in unified way, unlike other theories that appeal to different explanations to account for different cases. By comparing perceptual stimulus strength to mental strength, I showed that the tools developed by Signal Detection Theory provide a solid theoretical scaffolding for modeling variations in introspective accuracy and introspective confidence.

Chapter 3

Perception, Performance, and the Neural Correlates of Consciousness[†]

In the search of neural correlates of consciousness, subjects' response to the presentation of a visual stimulus can be assessed by subjective or objective measures (Snodgrass and Shevrin 2006; Seth et al. 2008; Sandberg et al. 2010; Irvine 2013). Researchers who use subjective reports as measures of the state of awareness of subjects recognize the importance of controlling for confounding factors (Dehaene and Changeux 2011; Merikle, Smilek, and Eastwood 2001; Sergent et al. 2013; Li, Hill, and He 2014; Bachmann 2009; Bachmann 2015). Ideally, when comparing a condition where subjects report consciously seeing a target against a control condition where subjects report not consciously seeing it, the difference between these two conditions should be conscious awareness only.

When looking for objective measures of conscious awareness, it is common that some researchers treat performance at chance level as a reliable indicator of unconscious processing (Dehaene et al. 1998; Eriksen 1960; Kouider and Dehaene 2007). The inability to distinguish a stimulus from noise or from another stimulus, however, should not be immediately equated with lack of awareness. Performance, at any level, should rather be treated as a potential confound in consciousness research (Dehaene and Changeux 2011; Lau 2008; Lau and Passingham 2006; Pitts et al. 2014; Li, Hill, and He 2014; Weiskrantz, Barbur,

---

[†] A version of this chapter was published as MORALES, J., Chiang, J., and Lau, H. (2015) "Controlling for Performance Capacity Confounds in Neuroimaging Studies of Conscious Awareness." *Neuroscience of Consciousness* 1(1). doi:10.1093/nc/niv008.

and Sahraie 1995; Aru, Bachmann, et al. 2012; Aru, Axmacher, et al. 2012; Bachmann 2009; Bachmann 2015). In contrast, subjective reports are indeed a valid measure of conscious awareness. As such, we should isolate the influence of task performance capacity in any comparison between different levels of subjective reports of awareness. However, even amongst those persuaded by this logic, few actually conduct experiments to isolate performance capacity confounds. The main reason is, probably, that it is difficult to achieve it experimentally. Usually, when subjective reports of awareness differ, performance capacity also differs. This is true in most detection and discrimination tasks, as well as in paradigms like binocular rivalry, in which detecting changes in the suppressed image is harder (Wales and Fox 1970).

Nevertheless, some attempts to control for performance capacity have been recently made in conscious awareness imaging studies. For example, Lau & Passingham (2006) conducted a study using metacontrast masking. By varying the SOAs between stimulus presentation and mask presentation, they found two SOAs where performance capacity in a discrimination task was matched for each subject, and yet subjective reports of awareness differed. They reported specific hemodynamic activation in the prefrontal cortex in association with trials in the condition that generated the higher percentage of "aware" ratings. This study can be taken as a proof of concept that performance capacity confounds can be eliminated. However, the number of trials where subjects claimed consciously seeing the target differed only by about 10% between the two conditions. Admittedly, a problem with this approach is that it relies on a specific kind of stimulus: metacontrast masked shapes. For researchers interested in other perceptual paradigms, it is hard to see how this method of performance capacity matching could generalize.

Another study (Persaud et al. 2011) matched performance between the normal sighted side of the visual field and the subjective blind side of the visual field in a hemianoptic patient, by presenting stimuli with low contrast to the patient's normal visual field and high contrast to the damaged visual field to compensate for the defects in processing sensitivity. But this opportunity is specific to the availability of a single rare patient.

While these studies effectively eliminated the performance confound as such, other problems intimately interlinked when controlling for performance can still arise. For instance, when performance is matched by varying the stimulation conditions, as in (Persaud et al. 2011), pre- and post-perceptual processing can obscure the interpretation of awareness-related activations (Bachmann 2009). Another potential issue is that subjective reports can differ due to variations in how subjects are probed and not due to differences in performance or conscious awareness itself. Different scales (Sandberg et al. 2011; Sandberg et al. 2010) or different criterion contents (i.e., different aspects of the experience subjects use for report) (Bachmann and Francis 2014; Bachmann 2015) can hinder contrastive analyses in imaging studies. Finally, another potential problem is that markers of specific conscious contents corresponding to the target stimulus have to be distinguished both conceptually and experimentally from the markers of conscious processes non-specific to the target. When attempting to eliminate performance confounds, this distinction is relevant because non-specific conscious processes can be shared by both correct and incorrect trials (Bachmann 2015). Unfortunately, it would be complicated to control experimentally for all these potential confounds at once (see Chapter 4 for more methodological issues).

In an attempt to overcome these difficulties, Lamy, Salti, & Bar-Haim (2009) proposed a general method to control for the influence of performance capacity by comparing between subjectively conscious and unconscious conditions during an EEG experiment. Instead of

trying to match performance experimentally, they proposed to correct for its influence mathematically, keeping stimuli at threshold constant across aware and unaware trials. In this chapter, I focus on this potentially promising method. First, we should expand on the logic of their methodology, trying to provide an intuitive explanation for the motivation behind it. Then, we should show that the method and its assumptions are problematic from the perspective of Signal Detection Theory (SDT) and offer an alternative based on it.

Although the focus is on Lamy and colleagues' proposal, it is important to note that it is useful to take it as a case study that has general conceptual and empirical ramifications concerning an appropriate analysis of perceptual signal, performance capacity confound, and the neural correlates of consciousness. Thus, the concerns raised regarding Lamy and colleagues' correction method can be generalized to other neuroimaging studies and techniques, as well as to philosophical debates on consciousness and its relation to performance in general and to attention in particular (Block 2007; Lau and Rosenthal 2011; Block 2010; Prinz 2012; Montemayor and Haladjian 2015). Furthermore, other laboratories have already used their suggested method (Hesselmann, Hebart, and Malach 2011) and leading consciousness researchers like Stanislas Dehaene have recently praised them for having accomplished the "remarkable feat" of keeping both performance and stimuli the same and, thanks to "a perfect control," having "confirmed [a neural] signature of conscious access" (Dehaene 2014, 129-30). However, despite all the merits behind it, their correction method makes unsound assumptions about perception and consciousness. Hence, its limitations have to be considered when designing and analyzing imagining studies on the neural correlates of consciousness.

1. Mathematical Correction for Performance Confound: Unconscious Lucky Answers

Lamy, Salti & Bar-Haim (2009) (LSB, henceforth) conducted an event-related potentials (ERPs) study on the neural correlates of conscious and unconscious visual processing where stimuli were constant across aware and unaware conditions. Subjects were presented with a 15x15 matrix of tilted lines (15˚), some of which were slightly more tilted (25˚) forming a 3x3 target square in one of four possible quadrants. A 15x15 matrix with tilted lines (25˚) masked the targets after a short (~25 to 100 ms, individually adjusted to achieve 25% conscious detection) or a long (~37 to 112 ms, individually adjusted to achieve 50% conscious detection) exposure. Subjects made two judgments. First, a 4-alternative forced choice (4-AFC) regarding the quadrant where the target 3x3 square was presented. Then, a subjective judgment whether they were aware of the target or whether they were just guessing. Continuous EEG (electroencephalography) was recorded from 20 scalp regions during all trials and subjects' responses were coded in the four following categories: subjects reported seeing the stimulus and correctly indicated its location (*aware-correct*), subjects reported seeing the stimulus and incorrectly indicated its location (*aware-incorrect*), subjects did not report seeing the stimulus and correctly indicated its location (*unaware-correct*) and subjects did not report seeing the stimulus and incorrectly indicated its location (*unaware-incorrect*). Note that in the last two categories subjects reported they were just guessing.

Confirming previous similar results (Batterink, Karns, and Neville 2012; Sergent, Baillet, and Dehaene 2005; Koivisto and Revonsuo 2010; Del Cul, Baillet, and Dehaene 2007), LSB reported a scalp-wide difference in the P3 waveform component (a positive voltage in the 300-650 latency range) in subjects' ERPs between the *aware-correct* and *unaware-correct*

conditions. They took this difference to reflect conscious processing. Critically, the comparison was focused on correct trials only (*aware-correct* vs. *unaware-correct*), as a direct comparison between all the *aware* and all the *unaware* trials would have involved a performance capacity confound. That is, awareness would have been confounded with overall performance since awareness co-occurred with higher performance rates. By comparing correct trials only, LSB matched performance in the sense that both conditions involve perfect accuracy, enabling thus a legitimate comparison between awareness and unawareness.

However, to really match performance between the conditions, LSB correctly realized the need to distinguish between two possible scenarios for trials in which subjects answered correctly and did not report seeing the target (the *unaware-correct* condition). It is possible that the subjects unconsciously processed the visual stimulus, and therefore answered correctly. Alternatively, the subjects could also have failed to process the stimulus, i.e., neither consciously nor unconsciously, and yet arrived at the correct answer by chance—in a 4-AFC task, random responding leads to an expected 25% chance of being correct. It is important to eliminate the influence of these correct-by-chance trials, because in comparing *aware-correct* and *unaware-correct*, the hope is not just to match performance as measured by sheer accuracy (in this case accuracy was 100% in both conditions). Rather, one would hope to match the underlying performance capacity. Only by removing the influence of the correct-by-chance trials in the *unaware-correct* condition one would be able to compare two conditions where the underlying performance capacities are matched (both at ceiling).

Thus, LSB developed a mathematical method to correct for the influence of those correct-by-chance trials (see Appendix and LSB's endnote 2). Their underlying idea is that by looking at the overall accuracy in *unaware* trials, one can estimate what percentage of trials in the

unaware-correct category is correct by chance. In a 4-AFC task we would expect 25% of unaware trials to be correct simply due to chance.

In order to correct for this percentage of unaware-correct-by-chance trials, LSB further assumed that the ERPs for these trials should just look like the ERPs for *unaware-incorrect* trials. The intuition behind their logic is that both types of trial have in common that subjects' brains failed to process the target. The only difference is that subjects were lucky in the correct-by-chance trials. With this assumption in mind, they attempted to subtract away the influence of the correct-by-chance trials on the set of unaware-correct trials. In summary, they assumed that the observed ERPs for overall *unaware-correct* is a weighted sum of the ERPs of the truly correct trials (processed-unaware-correct trials) and the ERPs of the correct-by-chance trials (unprocessed-unaware-correct trials). Thusly, their correction method would get at the underlying ERPs for the processed-unaware-correct trials, which they call *unaware-correct chance-free* trials (see Appendix and LSB's endnote 2 for details).

After this correction, LSB still found significant differences in the P3 components of ERPs between the *aware-correct* and the *unaware-correct chance-free* conditions. Because now both conditions were supposed to include only truly correct trials where the subjects processed the targets effectively, they argue that performance capacity was truly matched. Their logic is that their results now really reflect the signature of conscious processing, uncorrupted by confounds of performance capacity.

2. Problematic Assumptions of Mathematical Correction for Correct Trials by Chance

LSB analysis implicitly incorporates some of the major assumptions behind what is often called in psychophysics a High Threshold Model (HTM) (Swets 1961; Green and Swets 1966; Macmillan and Creelman 2005; Luce 1963). In this section, a general HTM is discussed in the context of detection and discrimination, and its discrepancies with the more popular methods of Signal Detection Theory (SDT).

a. High threshold models

A key conceptual component of HTM is that there is a discrete boundary that separates two distinct conditions: effective processing, in which a target is being processed correctly, and ineffective processing, in which a target is not being processed at all (Fig. 5). According to HTM, mere background noise can never lead to true detection, which means that correct responses during unprocessed trials arise only from guessing.

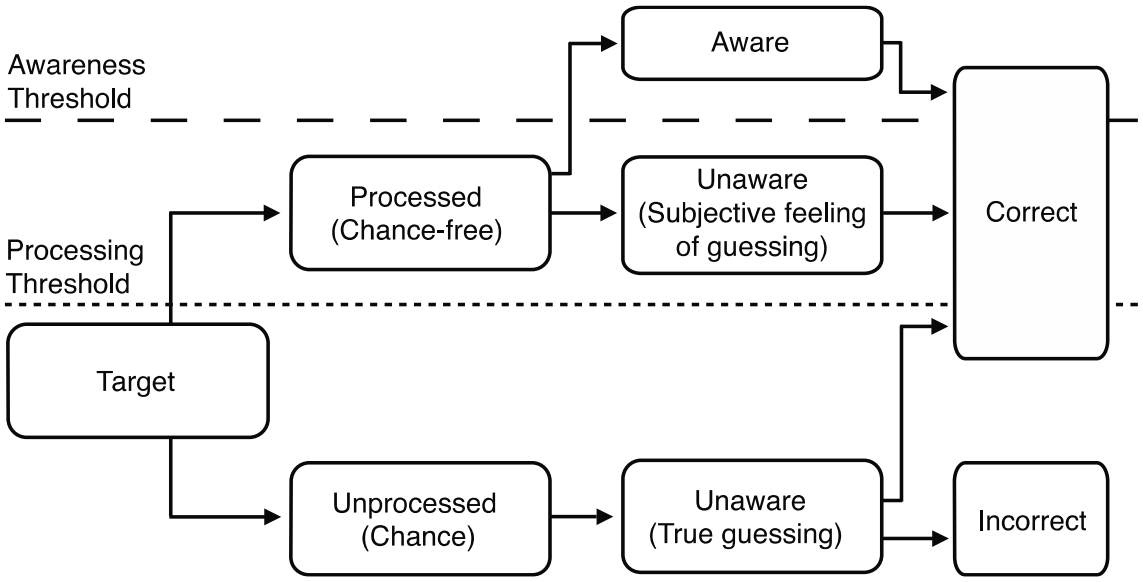

Figure 5. Schematic representation of LSB conceptual framework
Unprocessed targets lead to some correct responses due to luck. When the signal crosses the processing threshold (dotted line), the target is processed and it will always lead to a correct response (because the definition of 'processed' in this context means 'successfully processed'). Awareness requires crossing a further threshold (long dashes). Catch or lure trials (i.e., trials were there was no target presented) are left out from this schematic representation.

LSB seem to have in mind precisely this kind of model when discussing their experimental paradigm: "Because localization performance was clearly above chance, stimulus conditions were such that observers unconsciously perceived [i.e., processed] the target on average. Yet, on those individual trials in which the observers produced an incorrect response, *it is reasonable to claim that they did not perceive [i.e., processed] the target. Such trials were therefore defined as 'no-perception' trials*" (2009, 1442; emphasis added). Incorrect responses are a direct consequence, according to LSB, of a lack of processing of the target (bottom stream in Fig. 5) and, hence, of true guessing. LSB accept that perceptual processing is not sufficient for conscious awareness and, hence, that there can be processed unconscious targets (bottom half of top stream in Fig. 5). These trials are the ones that give rise to a

subjective feeling of guessing. Note that in their framework the unaware processed trials are always correct (because incorrect trials are no-perception trials). Put simply, for LSB only targets (i.e., never pure noise) can cross the processing threshold. Conversely, if a target is not reported accurately it can be inferred that it was not perceptually processed. The distinction between processed and unprocessed stimuli is, then, sharp and clear.

Following this model, the only possible source of ambiguity is those unprocessed (and hence unaware) responses that are correct due to chance (upward arrow in bottom stream on Fig. 5). LSB suggest comparing *unaware-correct chance-free* and *aware-correct* trials to find the true neural correlates of consciousness. In conclusion, the sharp distinctions between unaware-correct by chance, unaware-correct chance-free, and aware-correct trials that their proposal requires make sense only if something like HTM is assumed.

b. Signal detection theory

Despite its *prima facie* intuitiveness, decades of psychophysics research have favored Signal Detection Theory over High Threshold Models (Macmillan and Creelman 2005; S. A. Klein 2001; Luce 1963). Rather than having binary "processed" and "unprocessed" internal states, according to signal detection theory (SDT) the presentation of a target gives rise in the subject to an internal perceptual response that lies on a continuum (Fig. 6). The strength of the internal response is hardly ever exactly at zero due to the presence of noise. In other words, a stimulus is hardly ever in an unprocessed state. The signal of a target is always corrupted by noise, and therefore, performance capacity is determined by the signal-to-noise ratio of the internal response. There is no magical point below which subjects always completely fail to process the target and above which they always process it successfully.

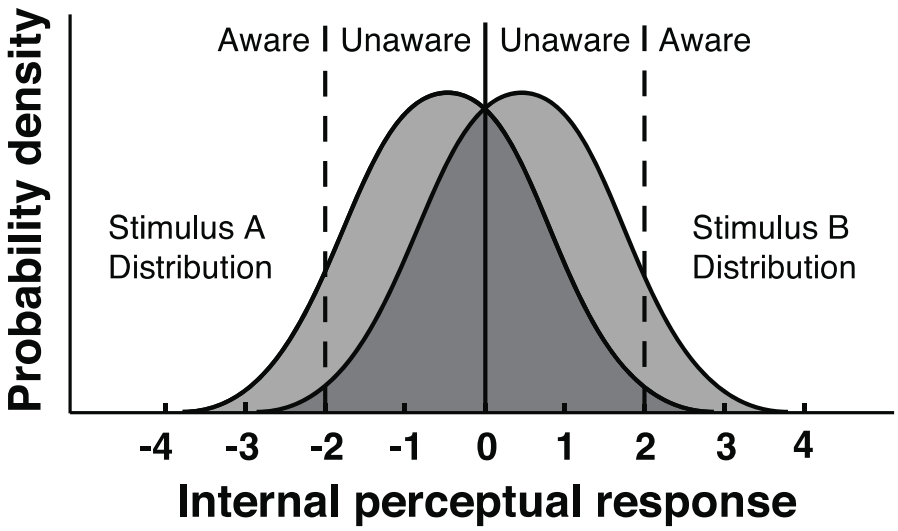

Figure 6. Signal detection theoretic model of perceptual awareness
The presentation of one of two stimuli evokes an internal response, falling into one of two Gaussian distributions. In each trial of a discrimination experiment subjects set a discrimination criterion (solid vertical line) and awareness criteria (dashed lines) against which they compare the internal response. Because the distributions overlap (darker area), it is possible (and quite common) that stimulus A is mistaken for stimulus B, or vice versa. Wrongly classified trials are reported as conscious if the internal response crosses the awareness criterion on the wrong side of the discrimination criterion.

According to SDT (Macmillan and Creelman 2005), the presentation of a stimulus A or B in a discrimination task gives rise to an internal response in the subject (Fig. 6). The internal perceptual response varies from trial to trial, falling into one of two Gaussian distributions with equal variance and different means, depending on the stimulus presented and the subject's internal state (i.e., noise). Subjects set a criterion against which they compare the internal response, which leads to the classification of the signal as being due to the presentation of stimulus A or B. The placement of the internal decision criterion is determined either by perceptual biases or by subjects' response biases (Witt et al. 2015). These can be influenced by preference, a strategy for maximizing the proportion of correct answers or expected value, subjective appearance (veridical or not) of the target, or

attentional resources (Macmillan and Creelman 2005; Rahnev et al. 2011; Morales et al. 2015). Because the distributions for the internal responses overlap, it is possible (and quite common) that stimulus A is mistaken for stimulus B, or vice versa. Additionally, trials are reported as *aware* when the internal perceptual response is strong enough to cross one of the outermost awareness criteria, and they are reported as *unaware* otherwise. Note that this allows for *aware-incorrect* trials when the internal response is drawn from the wrong distribution and yet it is strong enough to cross an awareness criterion (e.g., the right tail of the stimulus A distribution beyond the awareness criterion in Fig. 6).

Insofar as SDT rejects this strict dichotomy between perfectly processed and unprocessed stimuli, it is incompatible with HTM. But why prefer one model over the other?

c. The argument from incorrect conscious trials

A specific problem of HTM regarding consciousness studies is that it cannot explain the presence of incorrect trials when subjects report being aware of a target. According to the model as conceived by LSB, if subjects are aware of a target, it has to be because it was successfully processed. Thus, the presence of *aware-incorrect* trials is a problem. LSB report a small, but not negligible, percent of this kind of trials: 11% and 3.9% for short and long exposures, respectively. It is common practice in psychophysics to take into consideration lapse trials, i.e., trials where subjects did not witness the signal at all—sneezes or blinks are often blamed—or trials where non-perceptual problems, like motoric clumsiness, are accountable for the mistake. Lapse trials, however, are estimated at rates that go from 0% to 1% in the most lenient cases (Klein, 2001), which leaves LSB's empirical results unexplained.

However, *aware-incorrect* trials are not uncommon and they can be seen in many other studies (Hesselmann, Hebart, and Malach 2011), and in some cases in high proportions (Lau and Passingham 2006). Hence, the presence of *aware-incorrect* trials in LSB's experiment is in conflict with the core assumptions behind their version of a High Threshold Model. In contrast, as can be noted in Figure 6, *aware-incorrect* trials are an expected consequence of the SDT assumptions of the proposal advanced in this chapter. These trials are classified as aware and hence, despite being incorrect, should be accounted for when looking for the NCC.

d. Empirical inadequacy of HTM receiving operating characteristic curves

What really convinced generations of psychophysicists that SDT is a superior model to HTM is the comparison of theoretical and empirical ROC (receiver operating characteristic) curves. An ROC curve is a plot of hit rate against false alarm rate. In a discrimination task (but the principle generalizes to yes/no, detection, and forced-choice tasks as well), a subject's hit and false alarm rates produce one point on an ROC plot. By changing the subject's criterion in different conditions to be more liberal (more hits and more false alarms) and then more conservative (less hits and less false alarms), multiple points on the ROC space can be plotted. According to SDT, when sensitivity is different from zero, an ROC curve should be curvilinear (Fig. 7a), whereas according to HTM the ROC should be a straight line (Fig. 7b). Most empirical ROC curves from human subjects in visual experiments typically look like the one predicted by the SDT model, and hardly ever look like the one predicted by HTM (but see below). This is a strong reason to prefer SDT models over HTM with respect to human visual perception (Krantz 1969; Macmillan and Creelman 2005), auditory perception (Green and Swets 1966), and memory (Dube and Rotello 2012; Wixted 2009).

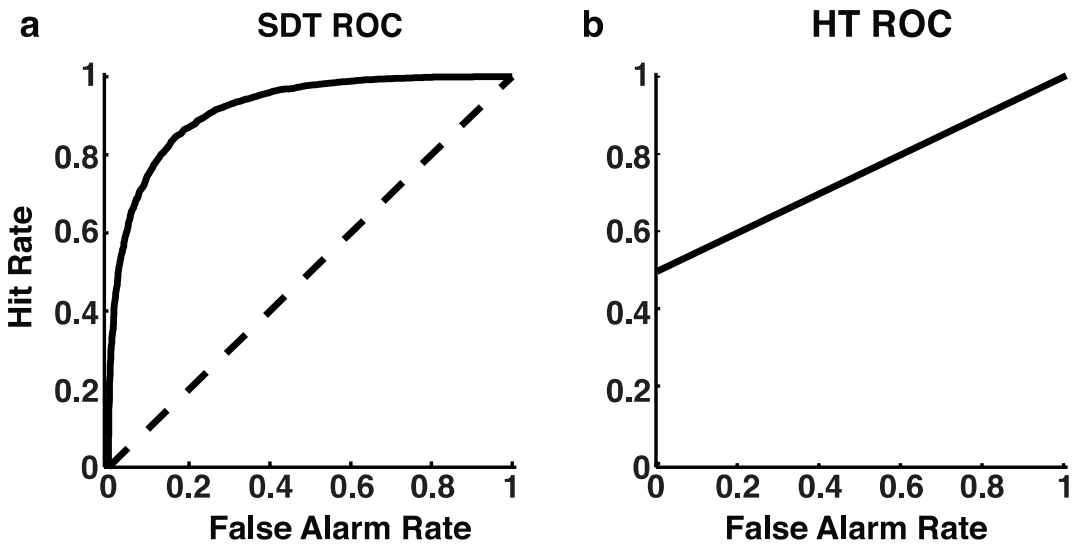

Figure 7. ROC curves comparison
(a) ROC curve as predicted by SDT. According to SDT, the tradeoff between having more hits and false alarms when there is non-zero sensitivity is a non-linear relationship determined by the signal-to-noise ratio. A zero-sensitivity scenario would yield a straight ROC line from zero to one (diagonal dashed line), where one can only increase hits by increasing the same amount of false alarms. A higher than zero signal-to-noise ratio means that the ROC curve will be curvilinear, where one can increase hits without increasing false alarms in the same proportion (solid curve) (i.e., performance above chance). ROC curve obtained from 10,000 simulated criteria for the same sensitivity level (d'=1). See Appendix for details regarding the simulation. (b) ROC curve as predicted by HTM. According to HTM, the vertical intercept is determined by the proportion of trials where the subject successfully processes the stimulus. The tradeoff between hits and false alarms follows a linear relationship.

It should be noted that in the memory literature, HTMs have enjoyed more popularity than in different perceptual modalities. In particular, mixed models (Aly and Yonelinas 2012; Yonelinas and Jacoby 2012), where recognition responses follow HTM and familiarity responses conform to SDT, have been well received, but they have also been criticized from the perspective of SDT (Wixted and Mickes 2010). Here we can be agnostic to this specific issue. The argument is not that all HTMs are necessarily wrong. What can be maintained here is that in the case of vision psychophysics, it is uncontroversial that SDT is much better

supported by empirical data than HTM and that HTMs are inappropriate for conscious awareness studies. Their inadequacy lies on how they depict the internal representation of signal and noise, heavily underestimating the role of the latter. Analysis methods for vision that assume HTM rather than SDT are, thus, problematic. But how problematic is LSB's High Threshold Model for conscious vision? How exactly might it have biased their results?

3. A Computer Simulation to Demonstrate the Inadequacy of the Correction Method

A computational simulation analysis was carried out to evaluate the degree of inadequacy of the correction method proposed by LSB. The idea behind it was to determine, assuming SDT is the correct model of perceptual processing (as the empirical evidence robustly suggests), how results of an idealized ERP experiment would look like using LSB's correction method. As any other theoretical model of perception, SDT has explanatory limits. It is only within these limits that the effectiveness of LSB's correction method was assessed.

For simplicity, it was assumed that subjects performed a 2-choice discrimination task, which is analytically more tractable than a 4-AFC task and its results are trivially generalizable. The simulation consisted on distinguishing between two stimulus alternatives (A & B), and then reporting whether there was awareness of the target or not. It followed the SDT assumptions presented in section 3.2. The presentation of a stimulus along with noise is assumed to give rise to an internal perceptual response that varies from trial to trial and that falls into one of two Gaussian distributions depending on which stimulus was presented. Discrimination is made by comparing the internal response to a criterion. The trial is

reported as aware if the strength of the internal response crosses one of the awareness criteria. For every trial, the strength of the internal perceptual response was correlated with a hypothetical neural response and a corresponding ERP of an arbitrary electrode site. This ERP was modeled as a sinusoidal response over time, scaling the amplitude of the ERP response by the strength of the internal perceptual response sampled from either of the Gaussian distributions (Fig. 8; see Appendix for technical details).

For computational simplicity, perceptual processing was modeled as the ERP response from 0 to 333ms. When the internal response was strong enough to cross the awareness criteria, the model assumes a constant brain signal is added to it, which may reflect a putative processing signature of awareness. For aware trials, then, an extra half cycle was added to the sinusoidal response so that there is a third "bump" in the ERP waveform (333ms to 500ms) (Fig. 8a). This extra cycle represents the differentiating processing uniquely associated with conscious awareness that is absent in trials without awareness (Figs. 8b & 8c). The idea is that by subtracting the *unaware* mean waveform from the *aware* mean waveform, if the unaware mean waveform is appropriately corrected for, we should be left just with activity properly related to awareness (i.e., the "third" bump). Despite its idealized nature, these simulations can help us determine the expected effectiveness of a performance correction method.

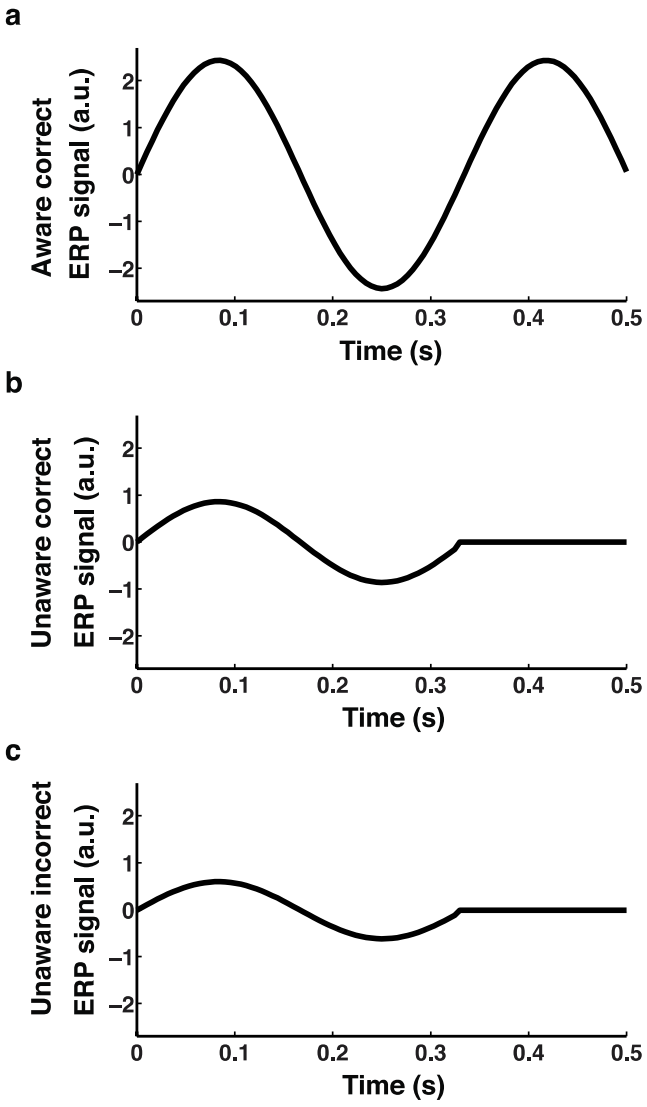

Figure 8. Average simulated waveforms from different conditions based on an SDT model

Representation of the mean ERP activation of three different conditions. There is an extra third "bump" in (a), the *aware-correct* trial, absent in (b) or (c), the unaware trials. This late activity is meant to reflect activity that is specific to awareness. Activity intensity in (c), the *unaware-incorrect* trial, is reduced compared to the higher activity in (b), the *unaware-correct* trial. (See Appendix for details.)

Neural responses associated with awareness need not arise late (>333ms) and they need not be temporally dissociated from the purely classification processes. Finding the precise timing and localization of these signatures is the goal of imagining studies looking for the

NCC. Hence, the simulations assumed the dissociated late timing for mere illustration purposes. The extra cycle associated with consciousness, then, could have been added earlier too (e.g., at ~100ms), as has been reported by different laboratories (Railo, Koivisto, and Revonsuo 2011; Rutiku et al. 2015; L. M. Andersen et al. 2015; Koivisto and Revonsuo 2003; Pins and ffytche 2003; Aru and Bachmann 2009). Along with other simplifications (e.g., the use of a sinusoidal waveform or the fact that wavelength, symmetry and latency are constant with changes in internal response), these assumptions should not affect the main lesson to be drawn from this exercise. Its main purpose is to illustrate how a correction method that assumes HTM performs under reasonable SDT assumptions. To emphasize, this simple model is suggested for ease of visualization and implementation only.

The results presented in Figures 8 and 9 were obtained after a 10,000-trial computer simulation (see Appendix for technical details). Figure 8 shows the ERP average responses under the different relevant conditions. In Figure 9, the correction was implemented as described by Lamy et al. (2009; specifically, endnote 2). The *unaware-correct* response (Fig. 8b, and repeated for ease of comparison in Fig. 9a as the solid curve) is only marginally different from the *unaware-correct chance-free* response (Fig. 9a, dashed curve). This is the waveform obtained after applying the correction suggested by LSB's method to eliminate performance confounds by lucky guesses. Hence, the influence of subtracting *unaware correct-by-chance* trials from *aware-correct* activations is only marginal. Both subtractive comparisons, namely, *aware-correct* minus *unaware-correct* (Fig. 9b, solid curve) and *aware-correct* minus *unaware-correct chance-free* (Fig. 9b, dashed curve), turn out to be almost the same, suggesting that the corrected unaware trials made a small contribution, if any, for singling out the signal specific to awareness. Concretely, in the latter comparison (Fig. 9b, dashed curve) there is still a clear residual activation during the first sinusoidal period of the

ERP (0–333ms), associated to the internal perceptual response strength in general, and not specifically to awareness, which occurs late in the simulations, i.e., from 333-500ms (Fig. 8a). An optimal analysis where only the awareness signature response remains after a subtractive comparison should cancel out the early response, leaving just the late response that is specific to awareness. As it is clear from Figure 9b, LSB's method fails to single out the specific response associated to awareness when plausible SDT assumptions are in place, defeating the purpose for which it was originally devised.

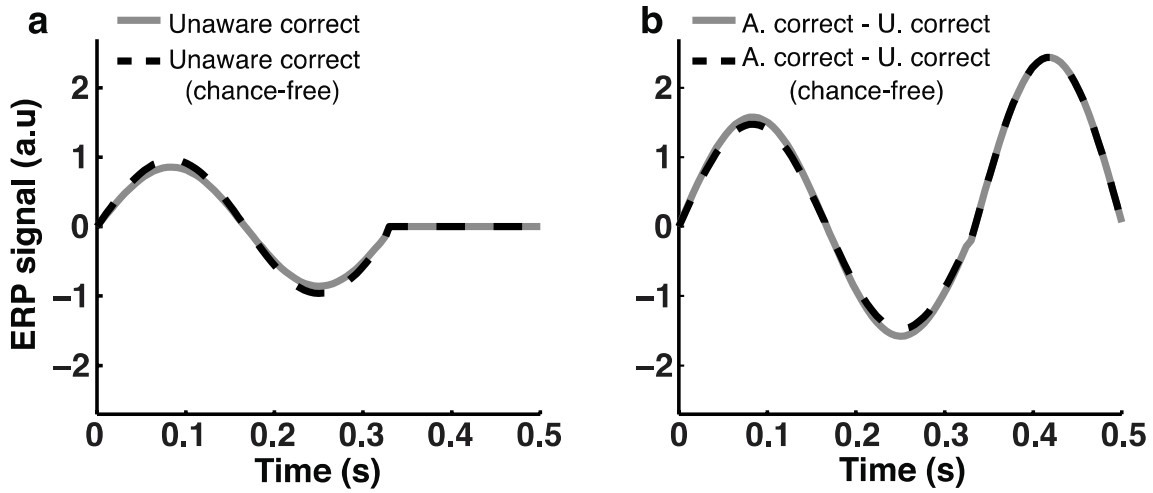

Figure 9. Simulated neural responses after LSB's correction method
(a) Simulated unaware-correct (solid curve) and unaware-correct chance-free (dashed curve) activation using LSB's (2009) suggested correction method. (b) The wave function result of subtracting the unaware-correct wave function (solid curve) and unaware-correct chance-free curves (dashed curve), respectively, from the awareness-correct wave function (Fig. 8a). It is evident from visual inspection that the influence of lucky responses was not sufficiently removed. Also, the activity during the early period was only marginally subtracted away with or without the correction method.

It is here that we can see the crucial, but flawed, role that LSB's High Threshold assumption plays. They assume that the ERP response of *unaware-correct-by-chance* trials

looks the same as the ERP response of *unaware-incorrect* trials (Fig. 8c): both are taken to be trials with unprocessed targets. Problematically, *unaware-incorrect* and *unaware-correct* do not look that different in the first place—the former's amplitude is only about half smaller than the latter's—so the *unaware-incorrect* waveform cannot have a very big influence on *unaware-correct* anyway. This is also observed in the actual ERP reported in LSB's 2009 paper (their Figure 3). It is of crucial importance to note that their results stayed basically the same regardless of whether they used *unaware-correct* or *unaware-correct chance-free* trials. In other words, their correction method affected in a negligible way their analyses, even though it was designed precisely to compensate for a significant underperformance during unawareness. This should be surprising for LSB since their assumed HTM implies that processed and unprocessed trials are radically different. Furthermore, in the P3 component during long-exposure trials (their Figure 2) there is no difference between the amplitude of *unaware-correct* and *unaware-incorrect* trials. This is an important unpredicted fact in their theory that receives no comment. (Note that the difference between *unaware-correct* and *unaware-incorrect* was found to be significant in the P3 component in the parietal region in a follow up study (Salti, Bar-Haim, and Lamy 2012)).

On SDT, however, this type of outcome is to be expected because both *unaware-correct* and *unaware-incorrect* are trials that come from the inner partitions between the awareness criteria, where signal strength is weak (Fig. 6), and they are not necessarily very different in each of the two partitions. As a matter of fact, *unaware-incorrect* trials may even have higher internal response strength than *unaware-correct* trials (due to the overlap of the Gaussian distributions), making them in the end qualitatively similar. In conclusion, LSB's correction method only partially, and inadequately, removes the performance capacity confound.

4. An SDT-based Correction Method

Having demonstrated the inadequacy of LSB's correction method, we can now show a way to perform a theoretically more adequate analysis based on SDT assumptions. The simulation presented in the previous section clearly established what the goal of such a correction should be, namely, to remove the ERP responses associated to mere processing in order to reveal the response that is specific to awareness and independent from performance. Like Lamy and colleagues, we are concerned with awareness as measured by subjective ratings (akin to confidence ratings as characterized within SDT). The distribution properties of the internal signal strength during a discrimination task are known when SDT is assumed, i.e., the internal perceptual response is drawn from one of two overlapping Gaussian distributions with equal variance and different means. Then, an appropriate correction for controlling for performance and factoring in any correct-by-chance trials is actually not difficult to achieve using standard SDT methods.

The primary assumption behind this correction is that activation intensity is linearly determined by the internal response. As it is clear from Figure 6, an SDT model assumes that unaware trials have a lower mean internal response than aware trials. This fact can be used to correct for performance confounds between aware and unaware trials. The ratio of the mean internal response for aware and unaware trials is used as a scaling factor of the unaware mean waveform. By scaling up the weaker response in the unaware condition to approximately match the intensity of the stronger response in the aware condition, we can subtract away any activation due to magnitude difference in internal response (see Appendix for technical details). Put simply, waveforms (but this is potentially generalizable to other types of imaging techniques like BOLD activity) of unaware trials during perceptual

processing must be scaled up to match waveform amplitudes (or activation) of aware trials before they are subtracted from them.

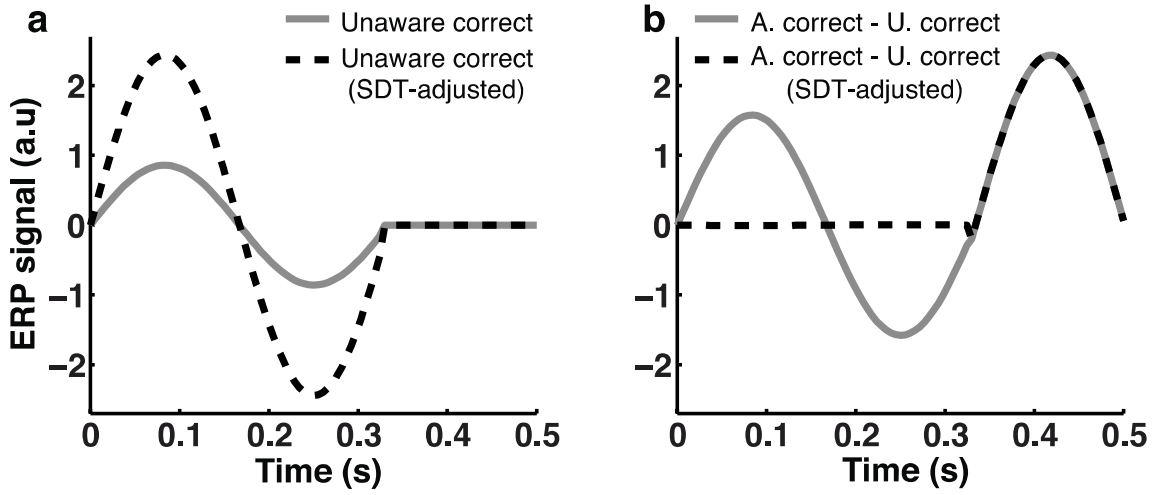

Figure 10. SDT-based correction method
(a) Simulated *unaware-correct* (solid curve) and *unaware-correct SDT-adjusted* (dashed curve) activation using the proposed SDT-based correction method (see Appendix for details). (b) *Aware-correct* activation curve corrected by subtracting *unaware-correct* wave function (solid curve; identical to solid curve in Fig. 9b, repeated here for ease of comparison) and *unaware-correct SDT-adjusted* curve (dashed curve), respectively. When comparing the corrected *aware-correct* curve in this figure to the one in Fig. 9b (dashed curve), it can be easily noticed by visual inspection that the proposed SDT-based adjustment method robustly removes the task performance capacity confound during the early processing stages, leaving just the awareness activation signature.

The correction from *unaware-correct* to *unaware-correct SDT-adjusted*, thus labeled to distinguish it from LSB's chance-free terminology, is presented in Figure 10a (dashed curve). The subtraction of the scaled up unaware waveform should leave us mainly with the activations relevant to awareness (i.e., the third "bump") in the simulated ERPs. Figure 10b shows the result of this process. For comparison, the subtraction *aware-correct* minus *unaware-correct* presented in Fig. 9b (solid curves) is repeated in 10b as well. Unlike LSB's

method, this adjustment method allows a significant difference between subtracting *unaware-correct* trials and *unaware-correct SDT-adjusted* trials.

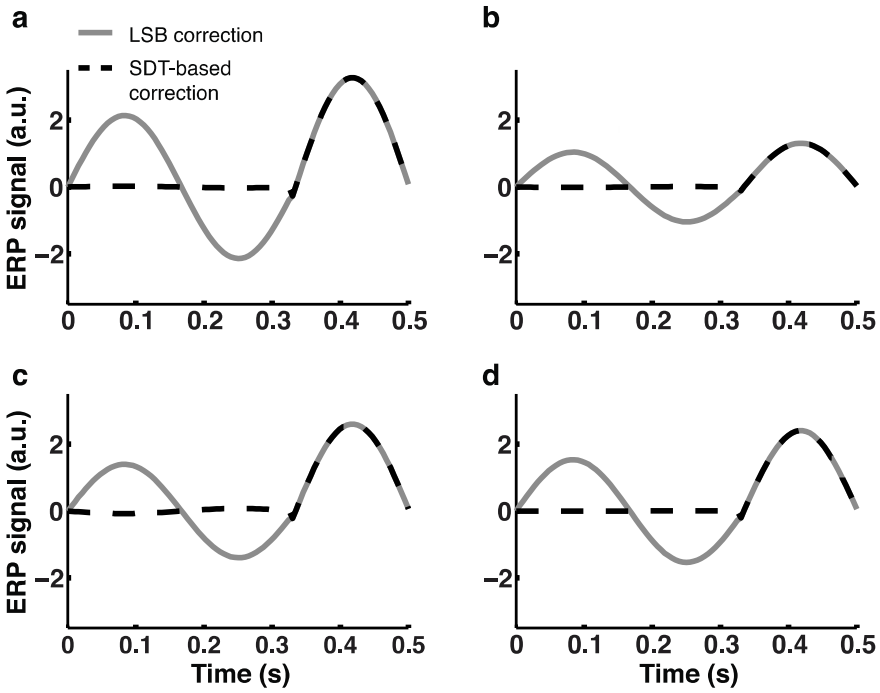

Figure 11. LSB's and SDT-based correction methods stimulation results under different parametric assumptions

10,000-trial simulations were performed, changing the parameters for awareness criteria and sensitivity (d'). The solid curve represents the subtraction of unaware-correct chance-free from aware-correct activation (i.e., LSB's method), and the dashed curve represents the subtraction of unaware-correct SDT-adjusted from aware-correct activation (i.e., the SDT-based correction method). (a) Sensitivity was kept constant and identical to previous simulations (d'=1). Very conservative awareness criteria were assigned, i.e., the internal response strength had to cross ±3 on the x-axis of Fig. 8a for a trial to be classified as aware. (b) Sensitivity was as in (a), but it was assigned a very liberal awareness criteria (±0.5). (c) Awareness criteria were held constant and identical to previous simulations (±2) and a higher sensitivity of d'=2 was assigned. (d) Awareness criteria were as in (c), but sensitivity was set to a low level of d'=0.5. The results for all four variations look qualitatively the same to the simulation results presented in Figures 9-10. See Appendix for details.

For the sake of completeness, Figure 11 includes the results of performing the same analysis with a different selection of parameters: better and worse performance (sensitivity

d') as well as more conservative and more liberal awareness criteria (see Fig. 11 caption and Appendix for details on the parameters used). Even though there is a slight numerical variation, changing simulated sensitivity or awareness criteria left intact the results thus far presented. The chance-free correction suggested by LSB is insufficient to isolate an awareness signature in the simulated ERP activation waveforms, while the SDT-based method is more robust to that end at the same time that it significantly reduces the worries regarding performance confound.

5. Conclusions

In order to discover the neural correlates of the exclusively subjective aspects of conscious awareness, eliminating performance capacity confound is a critical step. Lamy and colleagues' effort should be commended for recognizing the importance of this issue, and for providing a novel and general method for dealing with this problem in a formal way. The intuitive appeal of its core logic can be recognized as well as the importance and the potential impact that methods of its kind may have on the field. Unfortunately, whereas the overall concept behind the analysis is, *prima facie,* intuitive and appealing, it fails on a technical level due to its problematic assumptions.

The fact that the correction method proposed by LSB only minimally removes the performance capacity confound once plausible signal detection theoretic assumptions are made means that results based on it or on similar approaches have to be reassessed less optimistically. For instance, in their own study, LSB associated awareness with widespread activations. It would not be surprising that some of those activations are due to the failure to

thoroughly remove the performance capacity confound. Other laboratories (e.g., Hesselmann, Hebart, and Malach 2011) have used LSB's method trying to control for performance capacity and they found in an fMRI study that BOLD activity in the occipital and temporal areas was associated with awareness. But we know activity in some of these areas reflect internal response strength anyway (as they also predict task performance capacity), so their results may be merely due to the lack of complete removal of the influence of performance capacity. If this were the case, the view that these authors put forward, namely, that awareness may be associated with widely distributed activity in the whole brain, including visual areas, would be undermined. If an awareness signature response were correctly isolated, however, their findings may even turn out to be compatible with the view that awareness is associated with specific activity in a set of brain regions outside of the visual cortex, not directly involved in the generation of the internal perceptual response itself (Lau and Passingham 2006; Lau and Rosenthal 2011).

The simulation results are not presented without misgivings. They are highly idealized and they make strong parametric assumptions regarding neural data. For instance, they assume that the internal perceptual response follows strictly Gaussian distributions and that the strength of the ERP (or whatever other neural response is analyzed, like BOLD activity) follows the exact same distributions. We know that SDT models are appropriate for human perceptual behavior because the underlying parametric assumptions have been validated by psychophysical measurements of ROC curves, which show that the Gaussian distribution assumption is empirically justified in most cases of visual perception. Nevertheless, when it comes to ERP data, relatively little is known about their statistical nature. If awareness modulates neural activity nonlinearly (Friston et al. 1996), both the HTM and SDT

corrections presented in this chapter would fail to reveal the corresponding neural correlates properly.

Another limitation of the present work, shared by LSB's analyses, is that when contrasting unconscious and conscious activations, the latter could be revealing more than just the neural correlates of consciousness. These could also indicate brain activity present during conscious trials but unrelated to consciousness per se, like post-perceptual processing, working memory, or response preparation (Pitts et al. 2014; Aru, Bachmann, et al. 2012; Li, Hill, and He 2014; Bachmann 2009).

Finally, another limitation is that only one awareness criterion was assumed. This was done mainly for the sake of simplicity and computational tractability and it should not suggest that awareness is an on-off step function. Future work could pursue the effectiveness of this method with multiple criteria, which may more realistically capture the nature of subjective ratings. (Note that with enough criteria, the suggested type of modeling would, in practice, approximate a truly continuous scale.) Relatedly, it may be argued that there are subtle differences between confidence ratings (commonly used in SDT contexts) and awareness judgments (Overgaard and Sandberg 2012). We can acknowledge there are potential differences, but within the framework of SDT these two have been given similar treatments, in that they are both subjective ratings that can be modeled as responses separated by criteria.

With these caveats in mind, the conceptual ideas behind the SDT model are useful for the study of consciousness in both behavioral and imaging studies. Because this model is based on the localization of criteria along a decision axis, ratings of awareness can be dissociated from performance capacity, just as response bias can be dissociated from discrimination sensitivity (Maniscalco and Lau 2012). Furthermore, for a single trial, given

the internal response strength, the same stimulus could end up being classified as *aware* or *unaware* depending on where the criteria for awareness are placed. This is where HTM and SDT depart from each other more dramatically. Within SDT, for the same stimulus and the same internal response strength, the same subject could classify a trial as aware on one occasion and as unaware in a different occasion, depending on the localization of the subject's awareness criterion. This boundary is determined by fixating a criterion that changes from subject to subject, from experiment to experiment, and most likely it even jitters from trial to trial.

Perhaps, the most important take-home message of the exercise of focusing on LSB is not methodological in nature. Rather, there is a broader conceptual point that I am hoping to advocate here. When controlling for performance capacity in imaging studies, researchers should focus on controlling for the internal response strength, and not just for adjusting the influence of mere flukes. In imaging studies of consciousness, this means isolating some kind of further processing which only happens during trials crossing the awareness criteria. Such is the logic behind the proposed correction method. Given the complexity of this problem as revealed by the limitations of the correction method described here, in order to address the issue of performance capacity as a confound, the best method so far is to create task conditions in which task performance is empirically matched, and yet reported subjective levels of awareness differ (Lau and Passingham 2006; Rounis et al. 2010). Though this may be difficult to achieve experimentally, future research may be able to meet this important challenge.

Chapter 4

The Neural Correlates of Consciousness: Theories and Functions[†]

Our understanding of the neural basis of consciousness has substantially improved in the last few decades. New imaging and statistical techniques have been introduced, experiments have become more sophisticated, and several unsuccessful hypotheses have been quite conclusively ruled out. However, neuroscientists still do not entirely agree on the critical neural features required for sustaining perceptual conscious experiences in humans and other primates. In this chapter, a selection of influential views of the neural correlates of consciousness (NCC) and the predictions they make are discussed. For example, neural activity synchronized at 40Hz used to be considered a serious candidate for the NCC. Among current views, some expect activity in the ventral stream of the visual processing pathway to be crucial for consciousness, others expect recurrent activity in visual areas, distributed activity across frontoparietal areas, or specific activity in prefrontal cortex (PFC). In particular, the focus is placed on the predictions these views make with respect to the role of PFC during visual experiences, which is an area of critical interest and some source of contention. The discussion of these views will focus mainly on the level of functional anatomy, i.e., the level at which we consider different brain regions, rather than at the neuronal circuitry level. This approach makes sense because currently relatively more is understood about experimental evidence at this coarse level, and because these results are appropriate

---

[†] A version of this chapter is forthcoming as MORALES, J., and Lau, H. "The Neural Correlates of Consciousness." In U. Kriegel (Ed.) *Oxford Handbook of the Philosophy of Consciousness*. Oxford: Oxford University Press.

for arbitrating between current theoretical frameworks. For instance, while the Neural Synchrony Theory (Crick and Koch 1990), the Two-Visual-Systems Hypothesis (Milner and Goodale 2006), and the Local Recurrency Theory (Lamme 2010; Lamme 2006) predict that PFC activity is not critical for perceptual consciousness, the Higher Order (Lau 2008; Lau and Rosenthal 2011) and Global Workspace (Dehaene and Naccache 2001; Dehaene 2014; Baars 1988) Theories confer activity in PFC a crucial role in enabling conscious perception. Moreover, while Global Workspace Theory requires global and elevated activity distributed in a frontoparietal network, Higher Order Theory expects specific computations in PFC to be responsible for visual conscious experiences.

While it is sometimes described as a "brain mapping" issue (for example, in the form of questions like "*Where* is the neural basis of consciousness?"), finding the NCC is hardly a simple "localization" job. This is not to say that identifying certain areas differentially involved during conscious experiences is not part of what is required for finding the NCC. But the theoretically interesting quest for the NCC goes beyond straightforward "brain mapping." Success in finding the NCC is likely to involve describing how multiple brain areas work in conjunction to sustain conscious experiences, as well as the neural computations and the computational architecture behind them. Importantly, there are also important conceptual and experimental design issues that are relevant, where philosophy can play a key role. By highlighting some neurobiological and computational modelling results, it can be shown that the available evidence favors a hierarchical processing architecture that confers a crucial, if subtle and specific, role to PFC. After presenting the relevant results, methodological and functional implications of this neural architecture supporting conscious experiences are discussed. To anticipate, despite the apparent stark differences between conscious and unconscious perceptual processing, available evidence suggests that their

neural substrates must be largely shared. This indicates that the difference in neural activity between conscious and unconscious perceptual processing is likely to be subtle and highly specialized. In consequence, imaging techniques that focus only on marked differences between conscious and unconscious level of activity are likely to be insensitive to the relevant neural activity patterns that underlie conscious experiences. Finally, it follows from the evidence discussed here that the functional advantages of conscious over unconscious perceptual processing may be more limited than commonly thought.

1. Finding the Neural Correlates of Consciousness

Scientists study the neural difference between being conscious versus unconscious in at least two different ways. First, researchers may refer to the overall state of a person or animal (e.g., wakefulness, anesthesia, coma, sleep, etc.), for which they use phrases such as 'state-consciousness' or 'global states of consciousness'. Second, they may also refer to whether the person or animal is conscious of something or not (e.g., seeing or not seeing a face, seeing a face versus seeing a house, hearing or not hearing a sound, feeling or not feeling pain, etc.), also referred to as 'content-consciousness' or 'local states of consciousness'.[51]

When studying the NCC, scientists seek necessary and sufficient neural events that cause conscious experiences.[52] However, it has been acutely pointed out that finding necessary conditions for consciousness can be challenging (Chalmers 2000). First, after

[51] See note 28.

[52] The term 'correlate' falls short from capturing necessary and sufficient conditions. We just follow the terminology used in the field at least since (Crick and Koch 1990).

damage to a specific part of the brain (e.g., stroke, surgery, etc.), mental functions—including consciousness—may be lost. But they may also be recovered thanks to neuroplasticity: the brain's capacity to "rewire" itself. In some rare cases, cognitive functions and consciousness are never lost at all, even after massive, albeit slow, destruction of neural tissue (Feuillet, Dufour, and Pelletier 2007).

Second, redundancy makes finding necessary conditions for consciousness unlikely. It is not uncommon that the brain has redundant or backup mechanisms for performing the same function. This means that consciousness could be sustained by more than one neural mechanism. If mechanism *x* causally sustains consciousness, *x* is undoubtedly an NCC. But consciousness may be overdetermined if mechanisms *x* and *y* can cause the same type of conscious event independently. In this case, if *x* is damaged but *y* is spared, consciousness would still take place. This would demonstrate that *x* is not a necessary condition for that type of conscious event, even though it is *ex hypothesi* its neural correlate (or one of them). Thus, preservation of consciousness when a brain region is destroyed, impaired or when it does not display any measurable activity does not in and of itself show that normal activity in that region is not an NCC.

Third, convergent evolution could have produced independent mechanisms for consciousness in two species whose common ancestor lacked either mechanism. It may be the case that something as complex as consciousness emerged during evolution just once, but it is not necessary. If different species (say, humans and octopuses) sustain conscious experiences via different types of neural mechanisms, neither would be necessary for consciousness in a strong metaphysical sense. For all these reasons, establishing strict necessary conditions for consciousness is unlikely to be successful. If anything, we can aspire

to restricted necessity claims that include clauses like "in humans" or "in normal conditions".[53]

Finding sufficient neural conditions for consciousness is not without challenges either. For instance, everything else being the same, the whole brain is likely to be sufficient for sustaining conscious experiences. Yet, postulating the whole brain as the NCC would not be informative. Instead, neuroscientists are interested in the "*minimal* set of neural events jointly *sufficient* for a specific conscious experience (given the appropriate enabling conditions)" (Koch 2004, 97); or "core realizers" of consciousness for short (see Shoemaker 1981). Delimiting what counts as a core realizer is far from straightforward (Aru, Bachmann, et al. 2012; Chalmers 2000). For instance, when comparing a condition in which subjects report being conscious of a stimulus against a condition in which subjects report no consciousness of it, the difference between these two conditions should be conscious awareness only. Yet, distilling stimulation and cognition from consciousness is not easy. Controlling for stimulation, attention, and performance capacity (e.g., accuracy, reaction time, etc.), such that these are matched across conscious and unconscious conditions is hard to achieve experimentally (Lau 2008). During imaging experiments, prerequisites (e.g., stimulus processing, attention) and consequences (e.g., performance, attention, working memory, motor preparation, verbal report, etc.) of consciousness can be easily confounded with the actual NCC (Lumer and Rees 1999; Aru, Bachmann, et al. 2012; Bachmann 2015; Tse et al. 2005). Using lesion patients for whom performance is constant across subjective judgments of awareness and unawareness without experimental manipulation does not eliminate all the problems. Not only these patients are rare and their deficits are often

[53] Establishing what counts as a normal condition is complicated too, but we sidestep this issue here.

constrained in specific ways, their lesions are hardly ever limited to clear-cut anatomical or functional regions. Moreover, these patients' brains often rewire and recover functions in peculiar ways, which hinders making general inferences.

A practical limitation when studying the NCC is the methods currently available for detecting neural activity in the relevant functional networks. In the last few decades, sophisticated non-invasive imaging techniques such as functional magnetic resonance imaging (fMRI) have been added to decades-old technology like electroencephalography (EEG), magnetoencephalography (MEG), and positron emission tomography (PET). These technologies, however, have strong limitations with respect to either their spatial or temporal resolutions, or both. They are also indirect measurements of neural activity: oxygenated blood, electrical and magnetic signals measured outside the skull or glucose consumption detected via positron-emitting radioactive tracers. Electricocorticography (ECoG) allows making measurements with better signal-to-noise ratio and good temporal resolution by placing electrodes directly over the cortex, but it requires risky surgical intervention. For obvious medical and ethical reasons, the use of this technology in humans is very limited. In contrast, direct single- and multi-unit recording of neural activity offers unsurpassable spatiotemporal resolution. Unfortunately, it requires inserting electrodes directly into or right next to neurons, making it an extremely invasive method. In consequence, it is available almost exclusively in other animals like monkeys or rats. Working with animal models offers multiple advantages (Passingham 2009), but the study of consciousness may be challenging even when ingenious solutions have been devised (Leopold and Logothetis 1996; Rigotti et al. 2013). I will come back to some of the limitations of these methods when assessing the available empirical evidence for the NCC.

Finally, restricted necessary and sufficient conditions should ideally be established via causal interventions. By directly manipulating neural activity, we may reveal the causal mechanisms underlying conscious states (Craver 2007; Neisser 2012). Manipulating the brain safely and effectively, however, is a major challenge—especially in humans. Genetic, chemical, and surgical interventions are risky, almost exclusively available in other animals and likely to affect more than just conscious awareness. More promising may be the use of non-invasive technology such as transcranial magnetic stimulation (TMS). TMS pulses project a small magnetic field onto the surface of the brain through a coil placed outside the skull. Depending on the number and frequency of pulses, the magnetic field can enhance or inhibit neural activity in the target region. This allows researchers to create reversible "virtual lesions" for short intervals and test whether the target region was subserving the function of interest, including conscious awareness. While promising, the precise mechanisms of action of TMS are still poorly understood and its effects can only be coarsely controlled (Sandrini, Umiltá, and Rusconi 2011).

2. Theoretical Predictions Regarding the NCC

Different theories about the nature and localization of the NCC place their explanatory power at different levels (Hardcastle 2000). The emphasis has been sometimes laid on neurochemistry [e.g., activation of the NDMA neuroreceptor that forms large neural assemblies (Flohr 1995)], neuronal types [e.g., spindle neurons (Butti et al. 2013; Allman et al. 2005)], systemic properties [e.g., integrated information (Tononi 2008)], and functional neuroanatomy [e.g., specific neurophysiological markers and neural activity in specific regions or networks; for recent reviews see (Koch et al. 2016; Lau and Rosenthal 2011;

Dehaene and Changeux 2011)]. In this section, first, some important recent functional neuroanatomical theories are introduced. In no way this is an attempt at a thorough review. Not only other viable empirical theories of the NCC are not discussed, only succinct presentations of the ones discussed are offered. Rather, the goal is to show that the theories discussed here predict different neural implementations of consciousness, especially regarding the role of PFC, providing an opportunity to arbitrate empirically between several theoretical frameworks.

a. Neural synchrony theory

Much of the recent interest in finding the NCC was set off by the introduction of Neural Synchrony Theory (Crick and Koch 1990). According to it, at the psychological level consciousness depends on short-term memory and attention. At the neural level, attention makes groups of relevant neurons to fire in a coherent way giving rise to conscious percepts. Neurons in different areas often fire independently from each other. However, attention can make their firing rates to become synchronized in fast waves (between 40 and 70 times per second). This temporal coherence achieves a global unity imposed on different areas of the brain that activates short-term (working) memory. Crick and Koch hypothesize that this basic oscillatory mechanism underlies all kinds of consciousness (e.g., visual, auditory, tactile, or painful experiences). Thus, the NCC is identified in their theory with a special type of activity (i.e., neural firings oscillating at 40-70Hz). The specific contents of conscious experiences depend on the specialized cortex where the activity takes place. In the case of vision, different features of visual stimuli are processed by different areas of visual cortex (e.g., V1/orientation, V4/color, MT-V5/motion). The brain binds together all these features in

a single, coherent, and conscious percept by synchronizing the neural activity in these areas. Moreover, this activity is coordinated by zones in sensory cortices that are rich in feedback neurons (i.e., neurons that project from a higher area to a lower area). These feedback projecting zones also exist in other regions, such as the thalamus or the claustrum, which may play a major coordination role (Crick and Koch 2005). Thus, synchronized firing at about 40-70Hz is proposed as a necessary and sufficient condition for consciousness (provided enabling conditions such as attention and activation of working memory are met). Importantly, even though the NCC in Crick and Koch's proposal are highly distributed across brain areas, PFC is not predicted to play any significant role in sustaining conscious activity. At most, PFC may be relevant for attention, sustaining contents in working memory, and reporting conscious contents.

b. Two-visual-systems hypothesis

According to an influential theory advanced by Milner and Goodale (Milner and Goodale 2006), the neural correlates of visual awareness are restricted to activity in the ventral stream of the visual processing pathway. There are corticocortical projections from early visual cortex (V1) that later split into two processing streams (Ungerleider, Mishkin, and Mansfield 1982). One stream is located dorsally and ends in parietal cortex, the other stream runs on a ventral pathway that ends in inferior temporal cortex. The Two-Visual-Systems Hypothesis relies on neurophysiological and anatomical evidence in monkeys, as well as neuropsychological evidence in humans, to suggest activity in the dorsal stream is associated with visually-based action (for example, saccades or visually guided hand movements) and egocentric representations (i.e., representations of objects from the subject's point of view).

Despite involving complex computations, activity in this stream is not normally available to awareness according to this view. In contrast, activity in the ventral stream is typically associated with allocentric representations (i.e., objective representations independent of the subject's perspective) and visual object recognition. Objective visual representations have shape, size, color, lightness, and location constancies that allow subjects to re-identify objects independently of viewpoint (Burge 2010). Milner and Goodale argue that "visual phenomenology [...] can arise only from processing in the ventral stream." (Milner and Goodale 2006, 202) In other words, activity in the ventral stream is necessary for awareness. Additionally, attentional modulation that selects a represented object is required. Object representations in the ventral stream and attention are jointly sufficient for conscious awareness. Importantly, they think prefrontal cortex exert "some sort of top-down executive control [...] that can initiate the operation of attentional search" (Milner and Goodale 2006, 232), guide eye movements and motor control. However, activity in prefrontal cortex would probably be in and of itself irrelevant for conscious awareness.

c. Local recurrency theory

Local Recurrency Theory (LRT) proposes three stages involved in visual information processing. First, after stimulus presentation there is a rapid, unconscious feedforward sweep (~100-200ms) of activity from visual cortex (V1) to motor and prefrontal cortex. Immediately after, in a second processing stage, an exchange of information within and across high- and low-level visual areas starts taking place. This fast and widespread information exchange is achieved by means of so-called recurrent processing, namely, neural activity in horizontal connections within a visual area, and activity in feedback connections

from higher level areas back to lower levels (all the way back to V1). Local recurrent processing enables the exchange of information of different visual properties (e.g., orientation, shape, color, motion, etc.) that are processed independently in different visual areas. This facilitates the required "perceptual grouping" (Lamme 2006, 497) for forming coherent conscious representations of objects. According to LRT, this second stage of recurrent processing is the NCC as it is both necessary and sufficient for phenomenal consciousness (Lamme and Roelfsema 2000; Lamme, Zipser, and Spekreijse 2002).[54] Finally, in a late third stage, this reverberating activity becomes a widespread co-activated network involving visual and frontoparietal areas through attentional amplification. Motor and prefrontal cortex activity enables response preparation, keeping information in working memory and other types of cognitive control like attending, changing response strategies or inhibiting response. For LRT, this later frontoparietal activity is required exclusively for report and cognitive control (what Block (2007) calls 'access consciousness'), not for supporting conscious experiences themselves (what Block calls 'phenomenal consciousness'). One surprising consequence of the view is that conscious experiences take place even if they are not reportable or accessible to the subject (Block 2007; Landman, Spekreijse, and Lamme 2003; Sligte, Scholte, and Lamme 2008; Vandenbroucke et al. 2015). In other words, it would be possible to be conscious without knowing it and without any possible behavioral and cognitive manifestation of such phenomenal experiences.[55] In many cases, according to LRT,

---

[54] "That recurrent processing is necessary for visual awareness is now fairly well established, and supported by numerous experiments." (Lamme 2010, 216) "According to such empirical and theoretical arguments, [local recurrent processing] is the key neural ingredient of consciousness. We could even define consciousness as recurrent processing." (Lamme 2006, 499)

[55] See (Kouider, Sackur, and de Gardelle 2012; M. A. Cohen and Dennett 2011) for criticisms of the scientific viability of this position.

when subjects report unawareness, they may just be reporting their lack of access to otherwise conscious experiences.

d. Global workspace theory

According to Global Workspace Theory (GWT), after stimulus presentation, activity in visual areas starts accumulating in two independent processing streams, one that can lead to consciousness and another that supports unconscious processing (Del Cul et al. 2009; Charles et al. 2013; Charles, King, and Dehaene 2014).[56] Evidence accumulation through visual information processing in each stream races to a threshold in a "winner-takes-all" fashion (Shadlen and Kiani 2013; Pleskac and Busemeyer 2010; Wald 1947). If activity in the conscious stream reaches its threshold first, a sudden ignition "mobilizes" perceptual representations to a widespread global workspace implemented in frontoparietal interconnected neurons. This global broadcasting makes visual representations available for report and cognitive control, which results in a visual conscious experience (Dehaene and Changeux 2011; Dehaene and Naccache 2001). It is this globally broadcasted activity that GWT identifies as the NCC (Dehaene et al. 2006). Simultaneously, an unconscious stream processes the same visual stimulus. In case global ignition fails, the perceptual representation in the unconscious stream can be used if the subject is forced to provide a response, accounting for the commonly-observed capacity of subjects to perform above chance even when they are unaware of stimuli. Global workspace theorists appeal to a wealth of

[56] Not to be confused with the dorsal and ventral streams discussed by the Two-Visual-Systems Hypothesis. According to GWT, the conscious and unconscious streams may be implemented in largely overlapping anatomical regions in visual areas.

studies showing that all sorts of cognitive processing can be performed unconsciously to a certain extent: visual judgments, word meaning extraction, performing simple arithmetic operations, cognitive control, etc. (Dehaene et al. 2014). Note that this dual-stream approach makes the surprising assumption that every stimulus is processed twice simultaneously, which imposes stringent and possibly unnecessary computational requirements on the brain.

Global workspace theorists note that unconscious performance and neural activity associated to it are rarely at the same level as during conscious conditions. Thus, global ignition provides a necessary and minimally sufficient signature of consciousness, which according to the view, increases and maintains performance and cognitive flexibility. This signature is identified by GWT with frontoparietal activity in fMRI studies and with sudden, widespread activity in a late (~270-650ms) positive voltage in frontoparietal areas in EEG studies (also known as the P300 component) (Del Cul, Baillet, and Dehaene 2007; Sergent, Baillet, and Dehaene 2005; Lamy, Salti, and Bar-Haim 2009).

e. Higher order theory

The Higher Order Theory (HOT) of consciousness holds that a mental state is conscious by virtue of its relation to some higher-order state. A perceptual representation alone is never in and of itself conscious. Rather, it becomes conscious when it is somehow "tagged" or meta-represented by another, higher-order state. According to some versions of HOT, this relation is achieved by means of the higher-order state's representing the first-order state in ways similar to thought or perception (Rosenthal 2005). What different versions of higher order theories have in common is that "a mere change in the higher order representation or process

is sufficient to lead to a change in subjective awareness, even if all first-order representations remain the same" (Lau and Rosenthal 2011, 365).

HOT holds that first-order representations depend on neural activity in early visual areas, whereas higher-order processes are implemented mainly in prefrontal (and parietal) cortex in both human and other primates (Lau and Rosenthal 2011). More specifically, consciousness emerges from a hierarchical processing architecture in which unconscious visual information processed in early areas gets selected by downstream mechanisms in PFC. One of HOT's main predictions, then, is that disrupting the activity responsible for sustaining higher-order processes in prefrontal cortex should affect or eliminate visual experiences without affecting performance (because performance is driven mainly by unconscious first-order representations in early sensory cortex). Importantly, disruptions to PFC should affect conscious experiences themselves, not just report or access to visual experiences, as expected by LRT. In contrast to GWT, HOT does not expect global activity to be predictive of conscious awareness. PFC activity related to consciousness may be very subtle as it just needs to select relevant visual processes in early areas. Thus, HOT predicts that massive alterations to PFC may not be sufficient to disrupt consciousness as long as specific PFC activity is preserved. Perhaps more surprisingly, some versions of HOT predict that specific activity in PFC is necessary and minimally sufficient for consciousness. In other words, if the 'tagging' activity normally responsible for consciousness takes place in the absence of a 'tagged' state, conscious experiences may still occur.

In summary, these theories make very different general predictions about the nature and location of the NCC. They also make very different specific predictions regarding the role of PFC in consciousness, behavior, and the computational architecture underlying conscious

processing. Neural Synchrony Theory, Two Visual Systems Hypothesis, and Local Recurrency Theory focus on activity in sensory areas in fact denying any role in consciousness for PFC.[57] GWT accepts PFC plays an important role, emphasizing the heightened level of activity and its distribution through frontoparietal areas. In contrast, HOT confers PFC a dominant role in consciousness because of the specific and subtle function it plays within a hierarchical processing architecture.

A clear sign of progress in the scientific quest for the neural correlates of consciousness is that despite their initial popularity, some theories are completely abandoned in light of subsequent evidence. The Neural Synchrony Theory, for example, has lost credibility thanks to multiple studies finding oscillations at 40Hz in the absence of awareness and failing to detect these same oscillations during reports of conscious experiences (for a review, see Koch et al. 2016). The Two-Visual-Systems Hypothesis (at least with respect to its commitment to the ventral stream being the NCC) has also been subject of strong skepticism after considering the mounting evidence against the independence of the dorsal and ventral streams and their proposed clear-cut roles (Wu 2014a; Briscoe and Schwenkler 2015).

In the next two sections, I discuss neuroscientific and computational evidence relevant for arbitrating between the theoretical frameworks of the other three theories discussed in this section—LRT, GWT and HOT—and their predictions regarding the NCC and PFC's involvement.

---

[57] Neural Synchrony Theory and Local Recurrency Theory further specify that consciousness is associated with a specific type of feedback activity.

3. The NCC: Evidence of PFC's Involvement

Activity in PFC is crucial for supporting conscious perceptual experiences.[58] Multiple neuroimaging studies have systematically found increased activity in prefrontal and parietal cortex when comparing conscious versus unconscious conditions, often even when performance capacity is controlled for (Lau and Passingham 2006; Sergent, Baillet, and Dehaene 2005; Dehaene et al. 2001; for recent reviews, see Dehaene and Changeux 2011; Lau and Rosenthal 2011; Odegaard, Knight, and Lau 2017; Boly et al. 2017). Some researchers minimize PFC's importance in the NCC arguing that it plays an important function in attention, report, and cognitive control, but that it has a negligible role in consciousness (Koch et al. 2016; Tsuchiya et al. 2015). While these ideas are not new (Lumer and Rees 1999; Tse et al. 2005), they have sparked a renewed interest in the topic.

Admittedly, interpreting imaging results can be challenging. During an imaging experiment, reasons other than a causal role in supporting conscious experiences might lead to statistically significant results (e.g., noise or different functions performed by the same areas). As discussed in section 1, a more robust way of determining if an area of the brain is necessary for supporting a function is to permanently or temporarily impair it. If the function is lost, a constrained necessity claim may be warranted. Relatedly, if the function is not lost, not only constrained necessity claims are harder to maintain, the non-affected areas become candidates for being sufficient for supporting that function.[59] With this logic in mind, recent

---

[58] For simplicity we refer collectively to PFC, but activity relevant for consciousness is likely to be found in more specific areas, such as dorsolateral PFC, insula, and other orbitofrontal and rostrolateral regions.

[59] Necessity claims or denials in this context have to be constrained for the reasons discussed in the first section. Other species may implement consciousness differently, preventing any unconstrained

studies with carefully controlled psychophysical methods have investigated how PFC lesions (Del Cul et al. 2009; Fleming et al. 2014) and temporarily induced impairments by transcranial magnetic stimulation (Rounis et al. 2010) impact visual experiences. The results of these studies have been univocal: permanent and temporary impairments to PFC do not abolish objective visual task performance capacity, while they affect subjective judgments. Either the percentage of visible stimuli decreased despite constant performance (Rounis et al. 2010; Del Cul et al. 2009) or these subjective judgments became less diagnostic of task performance (Fleming et al. 2014). In the case of lesion patients, the capacity to use subjective ratings to diagnose task performance (i.e., metacognitive capacity) was impaired by 50% (Fleming et al. 2014).

Nevertheless, several objections are often raised against this evidence. First, it is argued that these impairments only affect subjective judgments mildly, while damage to early visual areas like V1 abolish visual consciousness completely; second, that PFC does not represent conscious content specifically, which confers it a limited role (if any); and, third, that the activity detected in PFC during imaging studies pertain to attention and report, not consciousness per se. These objections are addressed in order.

---

necessity claim. But, perhaps more importantly for the neuroscientific study of consciousness, failures to eliminate a function--consciousness in this case--need not imply that the area was not necessary (in a constrained way) for supporting the function. The impairment might not have been specific enough or the brain might have repurposed other circuits to implement that function which, otherwise, would have been implemented in the impaired area under normal conditions.

a. PFC activity related to consciousness is highly specific

Lesions to V1, in fact, can often completely abolish visual experiences (Melnick, Tadin, and Huxlin 2016; Weiskrantz 1986). When V1 is affected, like in blindsight, the sensory signal is degraded to the point of preventing subjective judgments of consciousness. In blindsight patients, the lateral geniculate nucleus (LGN) is spared. This relay center of visual information from the retina to early visual areas in the occipital lobe is located in the thalamus, and is likely responsible for driving objective performance of blindsight patients (Schmid et al. 2010). This does not rule out that in normal cases proper functioning of early visual areas is necessary, even if not sufficient, for consciousness.

A second point to highlight is that PFC functions very differently from sensory cortices. For instance, neuronal coding in PFC is relatively distributed, is rarely linear and shows a high degree of mixed selectivity (Mante et al. 2013; Rigotti et al. 2013). This means that, unlike visual cortex whose function is highly specialized for processing visual information, PFC's role in consciousness is performed by highly specific patterns of activity as it is responsible for carrying out many other functions as well. Therefore, to exclusively produce a large disruption of perceptual experience, neural patterns of activity in PFC would need to be affected in highly specific ways.

Relatedly, frontal and parietal cortices are densely connected and frontal regions display high neuroplasticity (Miller and Cohen 2001; R. A. Andersen, Asanuma, and Cowan 1985; Barbas and Mesulam 1981; Cavada and Goldman-Rakic 1989; Petrides and Pandya 1984; Croxson et al. 2005). This implies that the brains of patients with frontal impairments can rewire rapidly by the time they can be tested, often several months after the lesion. Lesions produced by trauma, stroke or ablation are often too unspecific, but sometimes they are

extended enough to likely include all regions responsible for consciousness. However, because these same regions support many central cognitive functions (Badre and D'Esposito 2009; Miller 2000; J. Duncan and Owen 2000; Passingham and Wise 2012), patients may be so generally impaired that testing them immediately following the brain damage may not be straightforward (Knight and Grabowecky 1995; Mettler 1949). As further support for this point, chemical inactivation in rodent and monkey PFC and regions strongly connected to PFC (e.g., pulvinar) lead to strong effects in subjective confidence judgments without affecting performance in perceptual and even memory tasks. In these cases, the animals are tested immediately after PFC or pulvinar are inactivated, preventing compensatory rewiring (Lak et al. 2014; Romanski et al. 1997; Shipp 2003; Komura et al. 2013; Pessoa and Adolphs 2010; Miyamoto et al. 2017). This background makes the specific effects of lesions or temporary impairments of PFC on subjective judgments indeed quite robust.

b. PFC encodes specific content

Another recent objection is that PFC activity does not encode specific content (Koch et al. 2016), making its role as the NCC likely to be limited. First, specific content representation of visual experiences in PFC is not explicitly predicted by all theories. For instance, PFC may enable conscious perception through connections to early visual areas where the specific content is supported (Lau and Rosenthal 2011). Second, and perhaps more importantly in terms of interpreting the available neuroscientific evidence correctly, denying that PFC represents explicit contents of conscious experiences is empirically unsupported.

Researchers often perform simple contrastive univariate analysis with fMRI data. In this kind of analysis, the overall levels of activity belonging to one experimental condition is

simply compared to (subtracted from) the overall levels of activity in another condition (e.g., conscious versus unconscious trials). But it is known that univariate fMRI analysis provides limited sensitivity. As mentioned above, activity in PFC is hardly linear and neurons exhibit mixed selectivity, which varies widely upon contextual changes. Measuring the overall levels of activity is at best a coarse approximation to total neural activity. Hence, visual content supported by specific patterns of activity may only be decoded effectively with careful analysis and sophisticated modelling strategies (Ester, Sprague, and Serences 2015; Stokes 2015). This includes multivariate analyses that go beyond a simple subtraction of overall activity. One example of this is multi-voxel pattern analysis (MVPA), where a decoder is trained to classify the *patterns* of activity in two conditions of interest. For example, if subjects are presented with two types of stimuli in different trials, say, houses and faces, the decoder can be trained to distinguish between patterns of activity pertaining to houses and patterns pertaining to faces. A successful decoder classifies above chance a novel set of data (usually data from the same subject that was not used during training) as belonging to house- or face-trials. MVPA reveals that perceptual content can be decoded from PFC in a simple perceptual decision task (Cortese et al. 2016), and that the pattern of activity in PFC reflects specific perceptual content even under several straining conditions (Wang, Arteaga, and He 2013). In another recent study, patterns of activity specific to subjective confidence judgments in perceptual and memory trials were successfully decoded from PFC (Morales, Lau, and Fleming 2018).

Finally, it could be objected that the spatiotemporal resolution of fMRI offers only a limited insight into neural activity, even when these sophisticated multivariate analyses are used. After all, it only gives us access to ~2 second snapshots of indirect blood-oxygen level dependent (BOLD) activity driven by the hundreds of thousands of neurons found in each

voxel (i.e., the minimum resolution in fMRI, equivalent to a 3D pixel of approximately 3 x 3 x 3mm). However, direct single- and multi-unit neural activity recording in monkeys offer a significantly higher spatiotemporal resolution (i.e., in the order of milliseconds and down to a single neuron) and multiple studies have unambiguously confirmed that specific perceptual decisions can be decoded from PFC (Kim and Shadlen 1999; Rigotti et al. 2013; Mante et al. 2013).

c. PFC is crucial for consciousness, not just attention or report

Together, the aforementioned evidence indicates that activity in PFC is necessary for visual consciousness. However, most of the fMRI studies mentioned above involved subjects explicitly reporting their conscious experience. A legitimate worry is that this activity does not reflect conscious perception *per se* and that, rather, it is confounded by the task demand to report or attend the stimulus (Koch et al. 2016; Tsuchiya et al. 2015). Some of these concerns have been recent rekindled by neuroimaging studies where subjects were not required to make explicit subjective judgments about visual stimuli and activity in prefrontal cortex previously related to consciousness was significantly diminished or undetected (Tsuchiya et al. 2015; Frassle et al. 2014).

The issues concerning limited sensitivity of methods commonly used in fMRI studies, specifically univariate analysis concerning PFC, are relevant here. Using more sensitive methods in humans, such as direct intracranial electrophysiological recording (electrocorticography, or ECoG), reveals activity related to visual consciousness in PFC even when subjects were not required to respond to the stimulus (Noy et al. 2015). Perhaps more importantly, in direct neuronal recordings in nonhuman primates who viewed stimuli

passively, activity specifically related to the stimulus was detected in PFC (Panagiotaropoulos et al. 2012). It could be argued, however, that even under passive viewing an over-trained animal may still attend the stimuli or implicitly prepare a report (which could increase prefrontal activity for reasons unrelated to consciousness). But even unreported features of a visual stimuli can be decoded from PFC activity. That is, even when the animal had to report on a different, orthogonal stimulus feature, the unattended and unreported feature was encoded in PFC (Mante et al. 2013). It is very unlikely that the monkeys prepared to attend or report on both features, especially considering that the task was challenging and involved near-threshold stimuli.

It is important to note that this does not mean that in studies of conscious perception making explicit reports does not further drive activity in PFC. PFC activity is involved in all sorts of higher cognition, not just conscious awareness. But this is consistent with the hypothesis that most univariate imaging techniques will only reveal the most heightened activity. It is also consistent with the observation by Noy and colleagues (2015) that their positive ECoG findings in PFC were subtle when no report was required. Still, in more direct recordings unreported stimulus features were robustly decoded, almost at the same level as attended and reported features (Mante et al. 2013). Thus, the objections from the so-called 'no-report' paradigms may have been exaggerated.

In summary, the important role of PFC in visual conscious experiences resists common objections. As anticipated in the first section, when looking for the NCC, methodological hurdles have to be considered with utmost care. When studying consciousness, non-invasive tools like fMRI may seem ideal for making inferences about neural function in humans. However, its spatiotemporal limitations as well as the prevalence of simple statistical

approaches should give us pause, especially when confronted with null findings. When ECoG and single- and multi-unit cell recordings along with multi-voxel pattern decoding analysis are incorporated, the picture that emerges is that activity in PFC is a serious candidate for being the NCC. This is incompatible with the main predictions made by LRT. Also, despite predicting an involvement of PFC during global ignition, GWT's requirement of global, heightened activity does not fit well with the evidence presented in this section. This evidence points towards a more subtle and specific role of frontal activity during conscious awareness. HOT also predicts an important role of PFC as the NCC but, in contrast to GWT, it does not require the relevant activity to be particularly heightened or distributed.

4. The Architecture of the NCC: Computational Considerations

Neuroimaging as well as direct cortical recordings offer evidence for determining where activity supporting conscious experiences is located in the brain. Multivariate analyses can even distinguish specific patterns of conscious and unconscious activity, rather than merely detecting a difference in levels of activity. Nevertheless, finding the NCC is not only a 'localization' problem. At the level of analysis that we are focusing on, it also involves finding the computational architecture most likely to account for the available neurophysiological and behavioral evidence. Computational modelling offers a non-invasive, formal way of comparing different models' capacities to account for behavioral data obtained in normal experimental conditions. Unlike neuroimaging and neurophysiology, where different conditions prevail across different experiments, in computational modelling the same data from a single experiment can be fed to a range of models. This is especially important for

comparing the likelihood of rival possible computational architectures of the NCC, giving them an equal chance to fit the data.

Some possible models of how perceptual processing and conscious processes interact in the brain are directly ruled out by the neurophysiological evidence. For example, a model that does not predict unconscious and conscious perceptual processing to take place in two distinct regions, like the one implied by LRT, is not particularly promising when evidence of the importance of frontal regions for visual consciousness is considered. Nevertheless, multiple computational architectures may be compatible with the extant neurophysiological evidence that privileges PFC. Unconscious and conscious processes could be instantiated in different fashions. For example, on one model these distinct processes could operate in parallel. On another model, perceptual conscious processing could operate hierarchically such that later activity associated with consciousness operates as if evaluating the quality of unconscious visual processes.

Let us explore this issue with the illustrative case of experiments in which performance is matched while subjective judgments differ. Humans and some nonhuman animals make perceptual decisions about the external world all the time, and they are also capable of making subjective judgments regarding the quantity, quality or reliability of their evidence regarding such perceptual decisions (e.g., by making one decision over another, by extending or suspending a search for resources, by providing visibility or confidence ratings, by placing bets regarding their likelihood of being correct, etc.) (Smith 2009; Beran et al. 2012; Fleming and Frith 2014).

Notoriously, objective perceptual decisions and subjective judgments about the stimuli can come apart in the laboratory and in clinical contexts. For instance, blindsight patients can objectively discriminate visual stimuli while denying having any subjective experience of

them (Weiskrantz 1986). In experimental conditions, humans (Maniscalco and Lau 2016; Rounis et al. 2010; Lau and Passingham 2006; Vlassova, Donkin, and Pearson 2014; Rahnev et al. 2011) and some other animals (Lak et al. 2014; Komura et al. 2013; Fetsch et al. 2014) can exhibit similar dissociations: subjects achieve comparable performance levels in a perceptual task while providing different subjective reports in different conditions. For example, in masking experiments (Maniscalco and Lau 2016; Del Cul et al. 2009; Lau and Passingham 2006), long and short gaps between stimulus presentation and the presentation of a mask allow subjects to identify the stimulus correctly at similar rates, while their subjective ratings of how visible the stimulus was differ significantly. These dissociations offer a unique opportunity to assess the specific processes involved in consciousness while distinguishing them from mere perceptual processing.

Let us consider three models recently used to fit data from a masking experiment (Maniscalco and Lau 2016): a single-channel, a dual-channel, and a hierarchical model (Figure 12). The single-channel model holds that subjective and objective judgments are different ways of evaluating the same underlying evidence generated by a single perceptual process. This sensory evidence consists on the sensory signal that arises in the brain after stimulus presentation plus the internal noise always present in neural processing. This sensory evidence is processed by the perceptual system and both objective and subjective systems tap into the same processing stream.

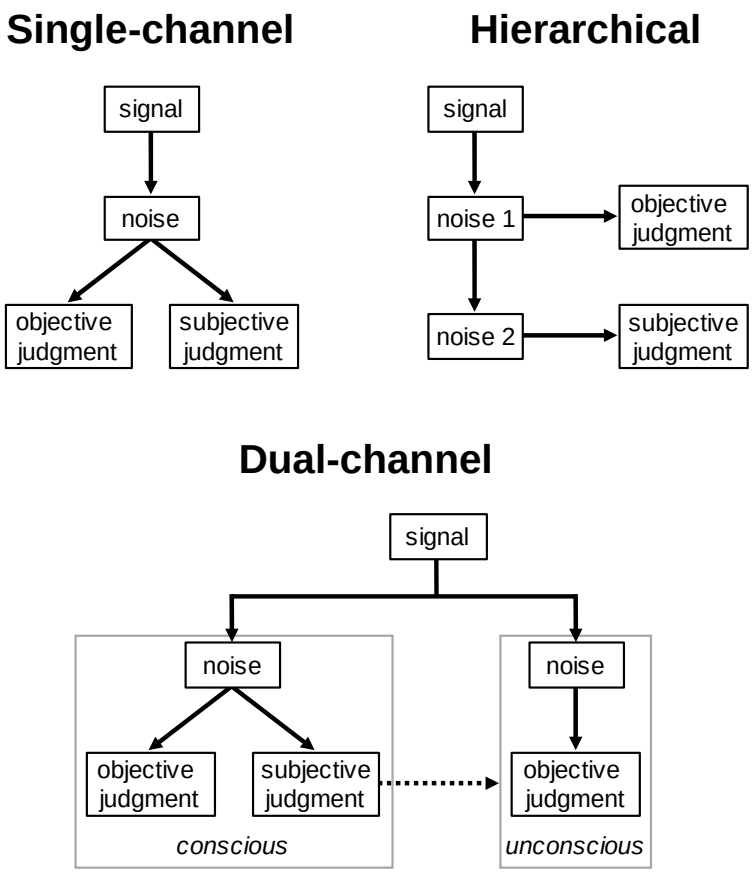

Figure 12. Diagrams of three computational models of objective and subjective judgments

Single-channel model. The same evidence (sensory signal + internal noise) gives rise to objective and subjective judgments. *Dual-channel model*. Two parallel streams of conscious and unconscious perceptual processing run simultaneously, each influenced by independent sources of noise. If the subjective judgment is given the lowest rating (e.g., "not seen", "not confident", "guess") the unconscious stream is used for objective classification, otherwise the conscious stream is used. *Hierarchical model*. Objective and subjective judgments are driven by different processes organized in a serial hierarchy. An early stage produces objective judgments and a later stage of processing produces subjective judgments, as if evaluating the quality of the former. The second stage inherits the noise of the first, influenced by the early stage, but not vice versa.

According to the dual-channel model, objective perceptual judgments are based on the same sensory evidence as subjective judgments when the subject is conscious of the stimulus, while unconscious perceptual judgments are based on an independent, parallel source of

evidence. 'Conscious' and 'unconscious' streams receive the same sensory signal but this gets affected independently by different sources of noise. If the conscious processing stream reaches a threshold first, the stimulus is classified by the brain as 'seen' and the sensory evidence is amplified and made available in working memory for further cognitive control (e.g., making a perceptual judgment about the stimulus and report that it was consciously seen). If the consciousness threshold is not crossed, the stimulus is classified by the brain as 'not seen' and the evidence accumulated in the conscious channel is discarded. If the subject still has to provide an answer—for instance, if prompted by the experimenter—the sensory evidence accumulated in the unconscious channel is used to provide a forced response.

Finally, according to the hierarchical model, the sensory evidence available for objective and subjective judgments differ, but it is not independent. The sensory signal (plus noise) is used to make objective perceptual judgments. Then, subsequent processing of this same evidence, in addition to a new source of noise, is used to make subjective judgments (Cleeremans, Timmermans, and Pasquali 2007; Fleming and Daw 2017). Thus, the accumulated evidence at the late stage might become degraded by the time it is tapped by subjective mechanisms due to signal decay or accrual of noise, or it may be enhanced due to further processing.

These models have been proposed based on conceptually reasonable grounds. In other words, finding which fits the empirical data better provides us with substantial insight regarding the computational architecture behind conscious perception. After performing formal model comparison, Maniscalco & Lau (2016) found that the hierarchical model provided the best and more parsimonious fit to the data of the metacontrast masking experiment, and it was also superior in reproducing the empirical data pattern in a series of simulations. The hierarchical model was able to account for the dissociation between

performance and subjective visibility ratings by supposing that early-stage perceptual processing is better transmitted to late-stage processing when the gap between stimulus and mask is longer. Since the early stage influences task performance and the late stage governs subjective judgments, longer gaps allow more evidence accumulation. This results in higher subjective visibility judgments in trials with longer gaps between the stimulus and the mask than in trials with short gaps, in spite of having similar task performance.

The last point is of importance for arbitrating between the theories of consciousness discussed in the previous sections. LRT does not make the prediction that the manipulation of the second processing stage changes subjective judgments without affecting task performance, bearing more functional resemblance to a single-channel model. Although GWT allows for unconscious above-chance performance, it does not predict unconscious performance will be at the same level if global workspace activity, likely implemented in frontoparietal regions, is disrupted. Some global workspace theorists explicitly endorse this dual-channel model which, at least for the masking dataset reported above, does not account well for the dissociation of objective and subjective judgments (Del Cul et al. 2009; Charles et al. 2013; Charles, King, and Dehaene 2014). The dual-channel model espoused by GWT, then, does not aptly account for the data presented in the previous section, where altering PFC normal functioning affects subjective judgments but preserves performance at normal levels (Rounis et al. 2010; Fleming et al. 2014). In any case, the idea of perfectly parallel processing routes for conscious and unconscious visual stimuli is unlikely to reflect the real neural circuitry involved in visual processing. As discussed above, for a long time the dorsal and ventral streams of visual processing were taken to be exclusively involved in unconscious and conscious visual representation, respectively (Milner and Goodale 2006). However, information within both streams is likely to be integrated (Wu 2014a), and unlikely to be

sufficient for consciousness. In sharp contrast to LRT and GWT, HOT predicts that late stage activity can be disrupted without affecting task performance. HOT explicitly proposes that downstream brain areas like PFC render sensory activity conscious by evaluating it. This puts HOT in close functional proximity to the hierarchical model, whose performance was far superior to the other two.

It is important to note that these results are limited to the analyzed dataset in Maniscalco & Lau (2016) and only further testing may confirm whether they generalize to other datasets, other experimental paradigms, or the hierarchical model outperforms other models. Nevertheless, it is also important to highlight that these results fit well with the data presented in the previous section according to which activity in PFC is crucial for conscious experiences. The second stage in the hierarchical model may be played by specific patterns of activity in PFC, while the earlier processing stage takes place in early visual areas.

5. Further Implications

The neuroscientific and computational evidence presented in the previous sections suggests that the NCC may be found in a hierarchical processing architecture of perceptual signals in the brain. In this section, let us explore some relevant implications of this architecture of the NCC.

a. Conscious and unconscious neural circuitry is largely shared

The Hierarchical model favored by the formal model comparison results holds that unconscious and conscious objective performance is based on the same perceptual evidence. Combined with available neuroscientific evidence, this suggests early visual and association areas support objective judgments while PFC taps onto this evidence later in the processing hierarchy, as if evaluating it, to give rise to consciousness. One consequence of this architecture is that, as far as visual information processing is concerned, unconscious and conscious mechanisms are mostly shared. PFC conscious-related engagement with visual representations constitutes only a late portion of the conscious processing stream, otherwise shared with unconscious representations. This important realization should impact how we study consciousness as well as how we think about the function of consciousness.

b. Distinguishing conscious and unconscious activity requires subtle methods

The fact that these mechanisms are largely shared points towards a subtle difference between conscious and unconscious processing. When controlling for stimulus strength and performance in an experimental setting, which is crucial for discovering the NCC, neural activity levels are not likely to differ greatly between conscious and unconscious trials. Activity in PFC is often not linearly correlated with behavior or stimulus properties and frontal neurons often have mixed selectivity properties that code distinct properties on a highly contextual manner (Rigotti et al. 2013; Mante et al. 2013). This suggests that we need to be very careful when interpreting results of purported elevated and distributed activity in conscious conditions (Lamy, Salti, and Bar-Haim 2009; Koivisto and Grassini 2016; Michel

2017; Railo, Koivisto, and Revonsuo 2011; Pitts et al. 2014; Dehaene 2014). In some of these experiments, it is often the case that stimulus strength and performance is inadequately controlled for and, sometimes, dated conceptions of the nature of perception hinder the interpretation of these results (see Chapter 3). For instance, it is easy to mistakenly include activity related to objective stimulus processing as part of activity responsible for consciousness.

The interpretation of null findings also demands caution. Detecting subtle neural activity specifically involved in consciousness requires highly sensitive methods. Current, non-invasive imaging technologies like univariate fMRI, MEG, or EEG are not ideal for such task as they are only able to detect the strongest signals from the brain. Because of their particular limitations and their indirect nature, subtle yet critical activity in prefrontal cortex is easily missed when comparing activity from conscious and unconscious conditions. In other words, while there may be nothing wrong with positive results when these methods detect strong activity in prefrontal cortex, we should be conservative about the meaning of null findings. The computational and empirical evidence gathered from more powerful methods suggest that, for the most part, only subtle and highly specific patterns of activity are relevant for consciousness. It should not be surprising then, that crude methods—advanced as they are—turn out to be often unsuited for detecting critical activity for consciousness in PFC.

c. The function of consciousness may be limited

If the mechanisms for unconscious and conscious processing are mostly shared and their difference is expected to be subtle and specific, it is possible that consciousness per se does

not contribute significantly to visual information processing, task performance or behavior in general (Rosenthal 2008; Robinson, Maley, and Piccinini 2015). It is hardly contested that the brain can perform lots of perceptual and cognitive tasks unconsciously (but see Phillips 2016; Peters and Lau 2015): anything from stimulus detection (Tsuchiya and Koch 2005) and word identification (Dehaene et al. 2001), to processing word meanings (Luck, Vogel, and Shapiro 1996; Gaillard et al. 2006) or performing basic arithmetic (Van Opstal, de Lange, and Dehaene 2011). Even high-level cognitive functions, like cognitive control (Koizumi, Maniscalco, and Lau 2015) or working memory (Samaha et al. 2016) show no apparent benefit from conscious awareness in controlled experimental conditions.

Denying the role of consciousness in behavior might strike as rather counterintuitive. Conscious experiences, it would seem, allow us to make fine-grained discriminations and to increase performance, and even to form beliefs, reason, and act (Tye 1996). In fact, in experiments showing above-chance performance in unconscious trials, the effects tend to be small and elicited only in forced-choice contexts. However, unconscious stimuli often differ from conscious ones in other ways besides consciousness. For instance, stimuli are often rendered unconscious by weakening perceptual stimulation (e.g., lower contrast, shorter presentation, higher noise, inattention, etc.), which has the effect of reducing the signal-to-noise ratio of the perceptual evidence. A lower signal-to-noise ratio alters first-order representations, expectedly decreasing performance capacity and the effect of attentional magnification. In these cases, it is the decreased signal-to-noise ratio elicited by the stimulation conditions rather than the stimulus being unconscious what accounts for the difference in performance capacity. This is why it is crucial to insist that performance capacity is a confound that needs to be controlled for when searching for the NCC (see Chapter 3).

This, of course, is not to deny consciousness has *some* function; although it does invite to rethink what the functions of consciousness might be. However, it does not seem to be a necessary trait of conscious experiences that they enable better performance than during unconscious processing. Still, some considerations are in place. First of all, it is worth emphasizing that in many experiments where unconscious performance is above chance it does not mean that it is as good as during conscious performance. Blindsighters, for instance, guess correctly very close to 100% of the time. But people with normal vision would respond correctly 100% of the time. This is consistent with the idea that when visual stimulation is matched, conscious vision might not yield any advantage over unconscious vision, but it might yield *some* advantage. This idea still goes against the often-assumed idea that consciousness offers a large advantage over unconscious processing.

As I argued in Chapter 1, it is quite possible that consciousness in general, and mental strength in particular, have a motivational function. It is hard to imagine how the initiation of directed actions would take place without consciousness. Blindsight patients, for instance, can detect and discriminate stimuli they are unaware of (Weiskrantz 1986). In some cases, they can even avoid obstacles while walking down a hallway (de Gelder et al. 2008). However, patients in these contexts are always prompted by the experimenters. In fact, they tend to be reluctant to comply with their instructions at first. The lack of awareness of the stimuli they are supposed to react to makes very unnatural for them to initiate behavior. Ned Block describes the imaginary case of superblindsighters (Block 1995). These are blindisighters who learn to prompt themselves to adventure guesses about objects in their blind fields. Eventually, their behavior would become indistinguishable from normal behavior despite the lack of consciousness. As far as I know, blindsighters cannot be thus trained. Finally, mental strength plays an important motivational role too. For example, there is less motivation to

take weak visual experiences at face value, and therefore act upon them, than when experiences are strong. Naturally, this important motivational role of mental strength is consistent with consciousness providing little to no advantage in visual processing.

6. Conclusions

The current science of consciousness is gradually achieving maturity. Fair assessments of empirical evidence related to the NCC, however, require subtle and thorough theoretical work. Determining necessary and sufficient neural conditions for consciousness goes beyond merely 'mapping' conscious-related activity (or lack thereof) onto certain brain areas. First, detecting or failing to detect activity in a brain area is not immediately uncontroversial evidence in favor or against that area being the NCC. For instance, activity in certain areas during conscious conditions may be confounded with activity of some other cognitive capacities related to performance, attention or cognitive control. Also, activity supporting consciousness in normal situations may be subtle and, hence, hard to detect with traditional methods. In consequence, scientists and philosophers need to be cautious as a few null results may not be sufficient for ruling out certain area as an important NCC. Second, a simple mapping of relevant brain areas is insufficient for explaining the overall computational architecture supporting consciousness. Even if certain brain area is found to be related to consciousness, activity in that area could be consistent with different processing architectures. So, the NCC is probably better understood as brain-wide interconnected processing rather than isolated activity in a single brain area.

Importantly, the empirical efforts behind the search for the NCC go beyond functional localization as they can also shed light on theoretical issues. As different theories make distinct predictions regarding the neurofunctional and computational architecture involved in consciousness, we can use empirical findings to arbitrate between these theories. The main predictions made by the Local Recurrency Theory regarding the NCC are not supported by current available evidence. A vast body of evidence using different methodologies privileges PFC as a crucial area for consciousness, which is incompatible with its central predictions. In contrast, both Global Workspace and Higher Order Theories predict PFC must have a major function in conscious awareness. From a study involving a formal model comparison, we concluded that a hierarchical computational model akin to HOT's prediction of a serial processing stream is better supported than a dual-channel model akin to some versions of GWT's prediction that objective and subjective processes are implemented in parallel. While this result is limited to the analyzed dataset, when considered along the systematic findings of PFC relevant role for consciousness, confidence in a hierarchical implementation of the NCC may be bolstered.

Finally, the data here presented point towards some important, although perhaps unexpected, features of the study of the NCC and consciousness itself. Firstly, the neural activity involved in conscious and unconscious perception may be largely shared. This suggests that the NCC involve subtle activity differences from unconscious processing which are detectable only by highly sensitive methods. Secondly, the function of consciousness may be limited. While a subtle difference in neural activity does not necessarily imply a subtle difference at the psychological, behavioral, or phenomenal level, it does make it a possibility. Only future research will be able to confirm or reject this hypothesis.

References

Allen, Colin. 1999. “Animal Concepts Revisited: The Use of Self-Monitoring as an Empirical Approach.” *Erkenntnis* 51 (1): 537–44. doi:10.1023/A:1005545425672.

Allen-Hermanson, Sean. 2015. “Introspection, Anton’s Syndrome, and Human Echolocation.” *Pacific Philosophical Quarterly* 98 (2): 171–92. doi:10.2307/1417725.

Allman, John M, Karli K Watson, Nicole A Tetreault, and Atiya Y Hakeem. 2005. “Intuition and Autism: A Possible Role for Von Economo Neurons.” *Trends in Cognitive Sciences* 9 (8): 367–73. doi:10.1016/j.tics.2005.06.008.

Alston, William. 1971. “Varieties of Privileged Access.” *American Philosophical Quarterly* 8 (3). North American Philosophical Publications: 223–41.

Aly, Mariam, and Andrew P Yonelinas. 2012. “Bridging Consciousness and Cognition in Memory and Perception: Evidence for Both State and Strength Processes.” Edited by Emmanuel Andreas Stamatakis. *PLoS ONE* 7 (1): e30231. doi:10.1371/journal.pone.0030231.s007.

Andersen, L M, M N Pedersen, K Sandberg, and M Overgaard. 2015. “Occipital MEG Activity in the Early Time Range (<300 Ms) Predicts Graded Changes in Perceptual Consciousness.” *Cerebral Cortex*, May. doi:10.1093/cercor/bhv108.

Andersen, R A, C Asanuma, and W M Cowan. 1985. “Callosal and Prefrontal Associational Projecting Cell Populations in Area 7A of the Macaque Monkey: A Study Using Retrogradely Transported Fluorescent Dyes.” *The Journal of Comparative Neurology* 232 (4): 443–55. doi:10.1002/cne.902320403.

Anton-Erxleben, Katharina, Christian Henrich, and Stefan Treue. 2007. "Attention Changes Perceived Size of Moving Visual Patterns." *Journal of Vision* 7 (11): 1–9. doi:10.1167/7.11.5.

Aristotle. 1994. *De Anima*. Translated by Jonathan Barnes. Oxford: Clarendon Press.

Armstrong, D M. 1968. *A Materialist Theory of the Mind*. London: Routledge & Kegan Paul.

Aru, Jaan, and Talis Bachmann. 2009. "Boosting Up Gamma-Band Oscillations Leaves Target-Stimulus in Masking Out of Awareness: Explaining an Apparent Paradox." *Neuroscience Letters* 450 (3): 351–55. doi:10.1016/j.neulet.2008.11.063.

Aru, Jaan, Nikolai Axmacher, Anne T A Do Lam, Juergen Fell, Christian E Elger, Wolf Singer, and Lucia Melloni. 2012. "Local Category-Specific Gamma Band Responses in the Visual Cortex Do Not Reflect Conscious Perception." *The Journal of Neuroscience* 32 (43): 14909–14. doi:10.1523/JNEUROSCI.2051-12.2012.

Aru, Jaan, Talis Bachmann, Wolf Singer, and Lucia Melloni. 2012. "Distilling the Neural Correlates of Consciousness." *Neuroscience and Biobehavioral Reviews* 36 (2): 737–46. doi:10.1016/j.neubiorev.2011.12.003.

Atlas, Lauren Y, Martin A Lindquist, Niall Bolger, and Tor D Wager. 2014. "Brain Mediators of the Effects of Noxious Heat on Pain." *Pain* 155 (8): 1632–48. doi:10.1016/j.pain.2014.05.015.

Aydede, Murat. 2009. "Is Feeling Pain the Perception of Something?" *The Journal of Philosophy* 106 (10). Journal of Philosophy: 531–67. doi:10.2307/20620203.

Aydede, Murat. 2017. "Pain: Perception or Introspection?" In *Routledge Handbook of Philosophy of Pain*, edited by Jennifer Corns. London: Routledge.

Baars, Bernard J. 1988. *A Cognitive Theory of Consciousness*. New York: Cambridge University Press.

Bachmann, Talis. 2009. “Finding ERP-Signatures of Target Awareness: Puzzle Persists Because of Experimental Co-Variation of the Objective and Subjective Variables.” *Consciousness and Cognition* 18 (3): 804–8. doi:10.1016/j.concog.2009.02.011.

Bachmann, Talis. 2015. “On the Brain-Imaging Markers of Neural Correlates of Consciousness.” *Frontiers in Psychology* 6 (June). doi:10.3389/fpsyg.2015.00868.

Bachmann, Talis, and Gregory Francis. 2014. *Visual Masking: Studying Perception, Attention, and Consciousness*. Oxford: Academic Press.

Badre, David, and Mark D’Esposito. 2009. “Is the Rostro-Caudal Axis of the Frontal Lobe Hierarchical?” *Nature Reviews Neuroscience* 10 (9): 659–69. doi:10.1038/nrn2667.

Bain, David. 2007. “The Location of Pains.” *Philosophical Papers* 36 (2): 171–205. doi:10.1080/05568640709485198.

Bantick, S J, R G Wise, A Ploghaus, S Clare, and S M Smith. 2002. “Imaging How Attention Modulates Pain in Humans Using Functional MRI.” *Brain* 125: 310–19.

Barbas, H, and M M Mesulam. 1981. “Organization of Afferent Input to Subdivisions of Area 8 in the Rhesus Monkey.” *The Journal of Comparative Neurology* 200 (3): 407–31. doi:10.1002/cne.902000309.

Batterink, Laura, Christina M Karns, and Helen Neville. 2012. “Dissociable Mechanisms Supporting Awareness: the P300 and Gamma in a Linguistic Attentional Blink Task.” *Cerebral Cortex* 22 (12): 2733–44. doi:10.1093/cercor/bhr346.

Bayne, Tim, and Maja Spener. 2010. “Introspective Humility.” *Philosophical Issues* 20 (1): 1–22.

Bayne, Tim, and Michelle Montague, eds. 2011. *Cognitive Phenomenology*. Oxford: Oxford University Press.

Bayne, Tim, Jakob Hohwy, and Adrian M Owen. 2016. “Are There Levels of Consciousness?” *Trends in Cognitive Sciences* 20 (6): 405–13. doi:10.1016/j.tics.2016.03.009.

Beck, J, and K Schneider. 2017. "Attention and Mental Primer." *Mind Language* 32 (4): 463–94. doi:10.1111/mila.12148.

Bennett, Jonathan. 1971. *Locke, Berkeley, Hume*. Oxford: Clarendon Press.

Beran, Michael J, Johannes Brandl, Josef Perner, and Joëlle Proust, eds. 2012. *Foundations of Metacognition*. Oxford University Press.

Block, Ned. 1995. "On a Confusion About a Function of Consciousness." *Behavioral and Brain Sciences* 18 (2): 227–47.

Block, Ned. 1996. "Mental Paint and Mental Latex." *Philosophical Issues* 7: 19–49. doi:10.2307/1522889.

Block, Ned. 2007. "Consciousness, Accessibility, and the Mesh Between Psychology and Neuroscience." *Behavioral and Brain Sciences* 30 (5-6): 481–548. doi:10.1017/S0140525X07002786.

Block, Ned. 2010. "Attention and Mental Paint." *Philosophical Issues* 20: 23–63.

Boghossian, Paul A, and J David Velleman. 1989. "Colour as a Secondary Quality." *Mind* 98 (389): 81–103. doi:10.2307/2255062.

Boly, Melanie, Marcello Massimini, Naotsugu Tsuchiya, Bradley R Postle, Christof Koch, and Giulio Tononi. 2017. "Are the Neural Correlates of Consciousness in the Front or in the Back of the Cerebral Cortex? Clinical and Neuroimaging Evidence." *The Journal of Neuroscience* 37 (40). Society for Neuroscience: 9603–13. doi:10.1523/JNEUROSCI.3218-16.2017.

Briscoe, Robert, and John Schwenkler. 2015. "Conscious Vision in Action." *Cognitive Science* 39 (7): 1435–67. doi:10.1111/mila.12056.

Burge, Tyler. 1996. "Our Entitlement to Self-Knowledge: I. Tyler Burge." *Proceedings of the Aristotelian Society* 96 (January). The Aristotelian Society: 91–116.

Burge, Tyler. 2010. *Origins of Objectivity*. New York: Oxford University Press.

Butti, Camilla, Micaela Santos, Neha Uppal, and Patrick R Hof. 2013. "Von Economo Neurons: Clinical and Evolutionary Perspectives." *Cortex* 49 (1): 312–26. doi:10.1016/j.cortex.2011.10.004.

Byrne, Alex. 2001. "Intentionalism Defended." *Philosophical Review* 110 (2). Duke University Press: 199–240. doi:10.2307/2693675.

Byrne, Alex. 2010. "Recollection, Perception, Imagination." *Philosophical Studies* 148 (1). Springer Netherlands: 15–26. doi:10.1007/s11098-010-9508-1.

Carrasco, Marisa. 2011. "Visual Attention: The Past 25 Years." *Vision Research* 51 (13): 1484–1525. doi:10.1016/j.visres.2011.04.012.

Carrasco, Marisa, Miguel Eckstein, Rich Krauzlis, and Preeti Verghese. 2013. "Attentional Modulation: Target Selection, Active Search and Cognitive Processing." *Vision Research* 85 (June): 1–4. doi:10.1016/j.visres.2013.05.001.

Carrasco, Marisa, Sam Ling, and Sarah Read. 2004. "Attention Alters Appearance." *Nature Neuroscience* 7 (3): 308–13. doi:10.1038/nn1194.

Carruthers, Peter. 2000. *Phenomenal Consciousness*. New York: Cambridge University Press.

Cavada, C, and P S Goldman-Rakic. 1989. "Posterior Parietal Cortex in Rhesus Monkey: II. Evidence for Segregated Corticocortical Networks Linking Sensory and Limbic Areas with the Frontal Lobe." *The Journal of Comparative Neurology* 287 (4): 422–45. doi:10.1002/cne.902870403.

Cervero, Fernando. 2012. *Understanding Pain: Exploring the Perception of Pain*. Cambridge, MA: MIT Press.

Chalmers, David. 2000. "What Is a Neural Correlate of Consciousness." In *Neural Correlates of Consciousness*, edited by Thomas Metzinger, 17–39. MIT Press.

Chalmers, David. 2003. "The Content and Epistemology of Phenomenal Belief." In *Consciousness: New Philosophical Perspectives*, 220–72. New York: Oxford University Press.

Chalmers, David J. 2010. *The Character of Consciousness*. New York: Oxford University Press.

Charles, L, J R King, and Stanislas Dehaene. 2014. "Decoding the Dynamics of Action, Intention, and Error Detection for Conscious and Subliminal Stimuli." *The Journal of Neuroscience* 34 (4): 1158–70. doi:10.1523/JNEUROSCI.2465-13.2014.

Charles, Lucie, Filip Van Opstal, Sébastien Marti, and Stanislas Dehaene. 2013. "Distinct Brain Mechanisms for Conscious Versus Subliminal Error Detection." *NeuroImage* 73 (June): 80–94. doi:10.1016/j.neuroimage.2013.01.054.

Chirimuuta, M. 2014. "Psychophysical Methods and the Evasion of Introspection." *Philosophy of Science* 81 (5): 914–26. doi:10.1086/677890.

Chudnoff, Elijah. 2015. *Cognitive Phenomenology*. Oxford: Routledge.

Cleeremans, Axel, Bert Timmermans, and Antoine Pasquali. 2007. "Consciousness and Metarepresentation: A Computational Sketch." *Neural Networks* 20 (9): 1032–39. doi:10.1016/j.neunet.2007.09.011.

Coghill, R C, C N Sang, J M Maisog, and M J Iadarola. 1999. "Pain Intensity Processing Within the Human Brain: A Bilateral, Distributed Mechanism." *Journal of Neurophysiology* 82 (4): 1934–43.

Coghill, Robert C, John G McHaffie, and Ye-Fen Yen. 2003. "Neural Correlates of Interindividual Differences in the Subjective Experience of Pain." *Proceedings of the National Academy of Sciences of the United States of America* 100 (14). National Academy of Sciences: 8538–42. doi:10.1073/pnas.1430684100.

Cohen, M R, and John H R Maunsell. 2009. “Attention Improves Performance Primarily by Reducing Interneuronal Correlations.” *Nature Neuroscience* 12 (12): 1594–1600. doi:10.1038/nn.2439.

Cohen, Michael A, and Daniel C Dennett. 2011. “Consciousness Cannot Be Separated from Function.” *Trends in Cognitive Sciences* 15 (8): 358–64. doi:10.1016/j.tics.2011.06.008.

Cohen, Michael A, Patrick Cavanagh, Marvin M Chun, and Ken Nakayama. 2012. “The Attentional Requirements of Consciousness.” *Trends in Cognitive Sciences* 16 (8): 411–17. doi:10.1016/j.tics.2012.06.013.

Cornoldi, C, R De Beni, F Giusberti, and E Marucci. 1991. “The Study of Vividness of Images.” In *Mental Images in Human Cognition*, edited by R H Logie. Elsevier.

Cortese, Aurelio, Kaoru Amano, Ai Koizumi, Mitsuo Kawato, and Hakwan Lau. 2016. “Multivoxel Neurofeedback Selectively Modulates Confidence Without Changing Perceptual Performance.” *Nature Communications* 7 (December): 13669. doi:10.1038/ncomms13669.

Coventry, Angela, and Uriah Kriegel. 2008. “Locke on Consciousness.” *History of Philosophy Quarterly* 25 (3): 221–42.

Crane, H D, and T P Piantanida. 1983. “On Seeing Reddish Green and Yellowish Blue.” *Science* 221 (4615): 1078–80. doi:10.1126/science.221.4615.1078.

Craver, Carl F. 2007. *Explaining the Brain*. Oxford University Press.

Crick, F C, and C Koch. 2005. “What Is the Function of the Claustrum?” *Philosophical Transactions of the Royal Society B: Biological Sciences* 360 (1458): 1271–79. doi:10.1002/cne.903350106.

Crick, Francis, and Christof Koch. 1990. “Towards a Neurobiological Theory of Consciousness.” *Seminars in the Neurosciences* 2: 263–75.

Croxson, Paula L, Heidi Johansen-Berg, Timothy E J Behrens, Matthew D Robson, Mark A Pinsk, Charles G Gross, Wolfgang Richter, Marlene C Richter, Sabine Kastner, and Matthew F S Rushworth. 2005. "Quantitative Investigation of Connections of the Prefrontal Cortex in the Human and Macaque Using Probabilistic Diffusion Tractography." *The Journal of Neuroscience* 25 (39). Society for Neuroscience: 8854–66. doi:10.1523/JNEUROSCI.1311-05.2005.

Cui, X, C B Jeter, D Yang, P R Montague, and D M Eagleman. 2007. "Vividness of Mental Imagery: Individual Variability Can Be Measured Objectively." *Vision Research* 47 (4): 474–78.

Cutter, B, and Michael Tye. 2011. "Tracking Representationalism and the Painfulness of Pain." *Philosophical Issues* 21: 90–109.

Dallenbach, Karl M. 1939. "Pain: History and Present Status." *The American Journal of Psychology* 52 (3). University of Illinois Press: 331–47. doi:10.2307/1416740.

Dauer, Francis W. 1999. "Force and Vivacity in the *Treatise* and the *Enquiry*." *Hume Studies* 25 (1-2): 83–99.

de Gelder, Beatrice, Marco Tamietto, Geert van Boxtel, Rainer Goebel, Arash Sahraie, Jan van den Stock, Bernard M C Stienen, Lawrence Weiskrantz, and Alan Pegna. 2008. "Intact Navigation Skills After Bilateral Loss of Striate Cortex." *Current Biology* 18 (24): R1128–29. doi:10.1016/j.cub.2008.11.002.

Dehaene, Stanislas. 2014. *Consciousness and the Brain*. New York: Viking Penguin.

Dehaene, Stanislas, and Jean-Pierre Changeux. 2011. "Experimental and Theoretical Approaches to Conscious Processing." *Neuron* 70 (2): 200–227. doi:10.1016/j.neuron.2011.03.018.

Dehaene, Stanislas, and Lionel Naccache. 2001. "Towards a Cognitive Neuroscience of Consciousness: Basic Evidence and a Workspace Framework." *Cognition* 79 (1-2): 1–37.

Dehaene, Stanislas, Jean-Pierre Changeux, Lionel Naccache, Jérôme Sackur, and Claire Sergent. 2006. "Conscious, Preconscious, and Subliminal Processing: a Testable Taxonomy." *Trends in Cognitive Sciences* 10 (5): 204–11. doi:10.1016/j.tics.2006.03.007.

Dehaene, Stanislas, L Naccache, G Le Clec'H, E Koechlin, M Mueller, G Dehaene-Lambertz, P F van de Moortele, and D Le Bihan. 1998. "Imaging Unconscious Semantic Priming." *Nature* 395 (6702): 597–600.

Dehaene, Stanislas, L Naccache, L Cohen, D L Bihan, J F Mangin, J B Poline, and D Rivière. 2001. "Cerebral Mechanisms of Word Masking and Unconscious Repetition Priming." *Nature Neuroscience* 4 (7): 752–58. doi:10.1038/89551.

Dehaene, Stanislas, Lucie Charles, Jean-Remi KING, and Sébastien Marti. 2014. "Toward a Computational Theory of Conscious Processing." *Current Opinion in Neurobiology* 25 (April): 76–84. doi:10.1016/j.conb.2013.12.005.

Del Cul, A, Stanislas Dehaene, P Reyes, E Bravo, and A Slachevsky. 2009. "Causal Role of Prefrontal Cortex in the Threshold for Access to Consciousness." *Brain* 132 (9): 2531–40. doi:10.1093/brain/awp111.

Del Cul, Antoine, Sylvain Baillet, and Stanislas Dehaene. 2007. "Brain Dynamics Underlying the Nonlinear Threshold for Access to Consciousness." *PLoS Biology* 5 (10): e260. doi:10.1371/journal.pbio.0050260.

Denison, Rachel. 2016. "Precision, Not Confidence, Describes the Uncertainty of Perceptual Experience." *Analytic Philosophy*, August, 1–22.

Dennett, Daniel C. 2002. "How Could I Be Wrong? How Wrong Could I Be?" *Journal of Consciousness Studies* 5-6: 13–16.

Descartes, René. 1985a. *Meditations on First Philosophy*. Edited by John Cottingham, Robert Stoothoff, and Dugald Murdoch. Vol. II. Cambridge: Cambridge University Press.

Descartes, René. 1985b. "Principles of Philosophy." In *The Philosophical Writings of Descartes*, edited by John Cottingham, Robert Stoothoff, and Dugald Murdoch. Vol. I. Cambridge: Cambridge University Press.

Descartes, René. 1985c. "The Passions of the Soul." In *The Philosophical Writings of Descartes*, edited by John Cottingham, Robert Stoothoff, and Dugald Murdoch. Vol. I. Cambridge: Cambridge University Press.

Descartes, René. 1985d. "Treatise on Man." In *The Philosophical Writings of Descartes*, edited by John Cottingham, Robert Stoothoff, and Dugald Murdoch. Vol. I. Cambridge: Cambridge University Press.

Desimone, Robert, and John Duncan. 1995. "Neural Mechanisms of Selective Visual Attention." *Annual Review of Neuroscience* 18: 193–222.

Donk, Mieke, and Wieske van Zoest. 2008. "Effects of Salience Are Short-Lived." *Psychological Science* 19 (7): 733–39. doi:10.1111/j.1467-9280.2008.02149.x.

Dretske, Fred. 1995. *Naturalizing the Mind*. Cambridge, MA: Bradford Books/MIT Press.

Dube, Chad, and Caren M Rotello. 2012. "Binary ROCs in Perception and Recognition Memory Are Curved." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 38 (1): 130–51. doi:10.1037/a0024957.

Dubin, Adrienne E, and Ardem Patapoutian. 2010. "Nociceptors: the Sensors of the Pain Pathway." *The Journal of Clinical Investigation* 120 (11): 3760–72. doi:10.1172/JCI42843.

Duncan, J, and A M Owen. 2000. "Common Regions of the Human Frontal Lobe Recruited by Diverse Cognitive Demands." *Trends in Neurosciences* 23 (10): 475–83.

Eriksen, Charles. 1960. “Discrimination and Learning Without Awareness: a Methodological Survey and Evaluation.” *Psychological Review* 67 (5): 279–300.

Ester, Edward F, Thomas C Sprague, and John T Serences. 2015. “Parietal and Frontal Cortex Encode Stimulus- Specific Mnemonic Representations During Visual Working Memory.” *Neuron* 87 (4): 893–905. doi:10.1016/j.neuron.2015.07.013.

Everson, Stephen. 1988. “The Difference Between Feeling and Thinking.” *Mind* 97 (387): 401–13. doi:10.2307/2255082.

Fazekas, Peter, and Morten Overgaard. 2017. “A Multi-Factor Account of Degrees of Awareness.” *Cognitive Science* 8 (10): 457. doi:10.1111/cogs.12478.

Fetsch, Christopher R, Roozbeh Kiani, William T Newsome, and Michael N Shadlen. 2014. “Effects of Cortical Microstimulation on Confidence in a Perceptual Decision.” *Neuron* 83 (4): 797–804. doi:10.1016/j.neuron.2014.07.011.

Feuillet, Lionel, Henry Dufour, and Jean Pelletier. 2007. “Brain of a White-Collar Worker.” *Lancet (London, England)* 370 (9583): 262. doi:10.1016/S0140-6736(07)61127-1.

Firestone, Chaz, and B J Scholl. 2015. “Cognition Does Not Affect Perception: Evaluating the Evidence for ‘Top-Down’ Effects.” *Behavioral and Brain Sciences* 39: 1. doi:10.1017/S0140525X15002691.

Fleming, Stephen M, and Christopher D Frith, eds. 2014. *The Cognitive Neuroscience of Metacognition*. Berlin: Springer.

Fleming, Stephen M, and Hakwan Lau. 2014. “How to Measure Metacognition.” *Frontiers in Human Neuroscience* 8 (July): 443. doi:10.3389/fnhum.2014.00443.

Fleming, Stephen M, and Nathaniel D Daw. 2017. “Self-Evaluation of Decision-Making: a General Bayesian Framework for Metacognitive Computation.” *Psychological Review* 124 (1): 91–114. doi:10.1037/rev0000045.

Fleming, Stephen M, Jihye Ryu, John G Golfinos, and Karen E Blackmon. 2014. "Domain-Specific Impairment in Metacognitive Accuracy Following Anterior Prefrontal Lesions." *Brain* 137 (10): 2811–22. doi:10.1093/brain/awu221.

Flohr, Hans. 1995. "Sensations and Brain Processes." *Behavioural Brain Research* 71 (1-2): 157–61. doi:10.1016/0166-4328(95)00033-X.

Forster, S, and N Lavie. 2016. "Establishing the Attention-Distractibility Trait." *Psychological Science* 27 (2): 203–12. doi:10.1177/0956797615617761.

Frassle, S, J Sommer, A Jansen, M Naber, and W Einhauser. 2014. "Binocular Rivalry: Frontal Activity Relates to Introspection and Action but Not to Perception." *The Journal of Neuroscience* 34 (5): 1738–47. doi:10.1523/JNEUROSCI.4403-13.2014.

Friston, K J, C J Price, P Fletcher, C Moore, R S Frackowiak, and R J Dolan. 1996. "The Trouble with Cognitive Subtraction." *NeuroImage* 4 (2): 97–104. doi:10.1006/nimg.1996.0033.

Fuller, S, Y Park, and M Carrasco. 2009. "Cue Contrast Modulates the Effects of Exogenous Attention on Appearance." *Vision Research* 49: 1825–37.

Fuller, Stuart, and Marisa Carrasco. 2006. "Exogenous Attention and Color Perception: Performance and Appearance of Saturation and Hue." *Vision Research* 46 (23): 4032–47. doi:10.1016/j.visres.2006.07.014.

Gaillard, Raphaël, Antoine Del Cul, Lionel Naccache, Fabien Vinckier, Laurent Cohen, and Stanislas Dehaene. 2006. "Nonconscious Semantic Processing of Emotional Words Modulates Conscious Access." *Proceedings of the National Academy of Sciences of the United States of America* 103 (19). National Academy of Sciences: 7524–29. doi:10.1073/pnas.0600584103.

Galton, Francis. 1880. "Statistics of Mental Imagery." *Mind* 5 (19): 301–18. doi:10.2307/2246391.

Gard, T, B K Holzel, A T Sack, H Hempel, S W Lazar, D Vaitl, and U Ott. 2012. "Pain Attenuation Through Mindfulness Is Associated with Decreased Cognitive Control and Increased Sensory Processing in the Brain." *Cerebral Cortex* 22 (11): 2692–2702. doi:10.1093/cercor/bhr352.

Garrett, Don. 2002. *Cognition and Commitment in Hume's Philosophy*. New York: Oxford University Press.

Gertler, Brie. 2001. "Introspecting Phenomenal States." *Philosophy and Phenomenological Research* 63 (2). International Phenomenological Society: 305–28.

Gertler, Brie. 2012. "Renewed Acquaintance." In *Introspection and Consciousness*, edited by Declan Smithies and Daniel Stoljar, 93–128. New York: Oxford University Press.

Gertler, Brie. 2018. "Self-Knowledge and Rational Agency: a Defense of Empiricism." *Philosophy and Phenomenological Research* 96 (1): 91–109. doi:10.2307/2108488.

Giustina, Anna, and Uriah Kriegel. 2017. "Fact-Introspection, Thing-Introspection, and Inner Awareness." *Review of Philosophy and Psychology* 8 (11): 143–64. doi:10.1007/s13164-016-0304-5.

Gobell, Joetta, and Marisa Carrasco. 2005. "Attention Alters the Appearance of Spatial Frequency and Gap Size." *Psychological Science* 16 (8): 644–51.

Goldman, Alvin. 2004. "Epistemology and the Evidential Status of Introspective Reports." *Journal of Consciousness Studies* 11 (7-8): 1–16.

Goldman, Alvin. 2006. *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. New York: Oxford University Press.

Gosseries, Olivia, Haibo Di, Steven Laureys, and Melanie Boly. 2014. "Measuring Consciousness in Severely Damaged Brains." *Annual Review of Neuroscience* 37 (1): 457–78. doi:10.1146/annurev-neuro-062012-170339.

Gracely, R H, R Dubner, and P A McGrath. 1979. “Narcotic Analgesia: Fentanyl Reduces the Intensity but Not the Unpleasantness of Painful Tooth Pulp Sensations.” *Science* 203 (4386): 1261–63.

Grahek, Nikola. 2007. *Feeling Pain and Being in Pain*. Cambridge, MA: MIT Press.

Green, David M, and John A Swets. 1966. *Signal Detection Theory and Psychophysics*. New York: John Wiley and Sons.

Hardcastle, Valerie Gray. 1999. *The Myth of Pain*. Cambridge, MA: MIT Press.

Hardcastle, Valerie Gray. 2000. “How to Understand the N in NCC.” In *Neural Correlates of Consciousness*, edited by Thomas Metzinger, 259–64. MIT Press.

Harman, Gilbert. 1990. “The Intrinsic Quality of Experience.” *Philosophical Perspectives* 4: 31–52.

Hatfield, Gary. 2005. “Introspective Evidence in Psychology.” In *Scientific Evidence: Philosophical Theories & Applications*, edited by Peter Achinstein, 259–86. Baltimore: Johns Hopkins University Press.

Hauck, Michael, J Lorenz, and A K Engel. 2007. “Attention to Painful Stimulation Enhances Γ-Band Activity and Synchronization in Human Sensorimotor Cortex.” *The Journal of Neuroscience* 27 (35): 9270–77. doi:10.1523/jneurosci.2283-07.2007.

Hesselmann, G, M Hebart, and R Malach. 2011. “Differential BOLD Activity Associated with Subjective and Objective Reports During ‘Blindsight’ in Normal Observers.” *The Journal of Neuroscience* 31 (36): 12936–44. doi:10.1523/JNEUROSCI.1556-11.2011.

Hobbes, Thomas. 1962. *Leviathan*. Edited by Michael Oakeshott. Oxford: Blackwell.

Hohwy, Jakob. 2009. “Consciousness and Cognition.” *Consciousness and Cognition* 18 (2): 428–38. doi:10.1016/j.concog.2009.02.006.

Hohwy, Jakob. 2011. “Phenomenal Variability and Introspective Reliability.” *Mind & Language* 26 (3): 261–86. doi:10.1016/j.tics.2006.05.002.

Hove, Michael J, Johannes Stelzer, Till Nierhaus, Sabrina D Thiel, Christopher Gundlach, Daniel S Margulies, Koene R A Van Dijk, Robert Turner, Peter E Keller, and Björn Merker. 2016. "Brain Network Reconfiguration and Perceptual Decoupling During an Absorptive State of Consciousness." *Cerebral Cortex* 26 (7): 3116–24. doi:10.1093/cercor/bhv137.

Hume, David. 2000. *A Treatise of Human Nature*. Edited by David Fate Norton and Mary J Norton. Oxford: Oxford University Press.

Irvine, Elizabeth. 2009. "Signal Detection Theory, the Exclusion Failure Paradigm and Weak Consciousness—Evidence for the Access/Phenomenal Distinction?" *Consciousness and Cognition* 18 (2): 551–60. doi:10.1016/j.concog.2008.11.002.

Irvine, Elizabeth. 2013. "Measures of Consciousness." *Philosophy Compass* 8 (3): 285–97. doi:10.1111/phc3.12016.

Itti, L, and C Koch. 2001. "Computational Modelling of Visual Attention." *Nature Reviews Neuroscience* 2 (3): 194–203. doi:10.1038/35058500.

James, William. 1950. *The Principles of Psychology*. New York: Dover.

Jiang, Y, P Costello, F Fang, M Huang, and S He. 2006. "A Gender- and Sexual Orientation-Dependent Spatial Attentional Effect of Invisible Images." *Proceedings of the National Academy of Sciences of the United States of America* 103 (45): 17048–52. doi:10.1073/pnas.0605678103.

Kepecs, Adam. 2013. "The Uncertainty of It All." *Nature Neuroscience* 16 (6): 660–62. doi:10.1038/nn.3416.

Kepecs, Adam, and Zachary F Mainen. 2012. "A Computational Framework for the Study of Confidence in Humans and Animals." *Philosophical Transactions of the Royal Society B: Biological Sciences* 367 (1594). The Royal Society: 1322–37. doi:10.1038/nature04676.

Kerzel, Dirk, Josef Schönhammer, Nicolas Burra, Sabine Born, and David Souto. 2011. "Saliency Changes Appearance." Edited by Marc O Ernst. *PLoS ONE* 6 (12). Public Library of Science: e28292–96. doi:10.1371/journal.pone.0028292.

Kiani, R, and M N Shadlen. 2009. "Representation of Confidence Associated with a Decision by Neurons in the Parietal Cortex." *Science* 324 (5928): 759–64. doi:10.1126/science.1169405.

Kim, J N, and M N Shadlen. 1999. "Neural Correlates of a Decision in the Dorsolateral Prefrontal Cortex of the Macaque." *Nature Neuroscience* 2 (2): 176–85. doi:10.1038/5739.

Kind, Amy. 2003. "What's So Transparent About Transparency?" *Philosophical Studies* 115 (3). Springer: 225–44. doi:10.2307/4321404.

Kind, Amy. 2017. "Imaginative Vividness." *Journal of the American Philosophical Association* 3 (1): 32–50. doi:10.1017/apa.2017.10.

Kingdom, Frederick A A, and Nicolaas Prins. 2010. *Psychophysics*. Cambridge, MA: Academic Press.

Klein, Colin. 2015. *What the Body Commands*. Cambridge, MA: MIT Press.

Klein, Colin, and Manolo Martínez. 2016. "Imperativism and Pain Intensity." Edited by David Bain, M Brady, and Jennifer Corns. *The Nature of Pain*.

Klein, Stanley A. 2001. "Measuring, Estimating, and Understanding the Psychometric Function: a Commentary." *Perception & Psychophysics* 63 (8): 1421–55.

Knight, Robert T, and Marcia Grabowecky. 1995. "Escape From Linear Time: Prefrontal Cortex and Conscious Experience." In *The Cognitive Neurosciences*, edited by Michael S Gazzaniga, 1357–71. Cambridge, MA: The MIT Press. http://doi.apa.org/psycinfo/1994-98810-090.

Koch, Christof. 2004. *The Quest for Consciousness*. Roberts Publishers.

Koch, Christof, Marcello Massimini, Melanie Boly, and Giulio Tononi. 2016. "Neural Correlates of Consciousness: Progress and Problems." *Nature Reviews: Neuroscience* 17 (5): 307–21. doi:10.1038/nrn.2016.22.

Koivisto, Mika, and Antti Revonsuo. 2003. "An ERP Study of Change Detection, Change Blindness, and Visual Awareness." *Psychophysiology* 40 (3): 423–29.

Koivisto, Mika, and Antti Revonsuo. 2010. "Event-Related Brain Potential Correlates of Visual Awareness." *Neuroscience and Biobehavioral Reviews* 34 (6): 922–34. doi:10.1016/j.neubiorev.2009.12.002.

Koivisto, Mika, and Simone Grassini. 2016. "Neuropsychologia." *Neuropsychologia* 84: 235–43. doi:10.1016/j.neuropsychologia.2016.02.024.

Koizumi, Ai, Brian Maniscalco, and Hakwan Lau. 2015. "Does Perceptual Confidence Facilitate Cognitive Control?" *Attention, Perception, & Psychophysics* 77 (4): 1295–1306. doi:10.3758/s13414-015-0843-3.

Komura, Yutaka, Akihiko Nikkuni, Noriko Hirashima, Teppei Uetake, and Aki Miyamoto. 2013. "Responses of Pulvinar Neurons Reflect a Subject's Confidence in Visual Categorization." *Nature Neuroscience* 16 (6): 749–55. doi:10.1038/nn.3393.

Kosslyn, Stephen M. 1996. *Image and Brain*. Cambridge, MA: Bradford Books.

Kouider, Sid, and Stanislas Dehaene. 2007. "Levels of Processing During Non-Conscious Perception: a Critical Review of Visual Masking." *Philosophical Transactions of the Royal Society B: Biological Sciences* 362 (1481): 857–75. doi:10.1098/rstb.2007.2093.

Kouider, Sid, Jérôme Sackur, and Vincent de Gardelle. 2012. "Do We Still Need Phenomenal Consciousness? Comment on Block." *Trends in Cognitive Sciences* 16 (3): 140–41. doi:10.1016/j.tics.2012.01.003.

Krantz, David H. 1969. "Threshold Theories of Signal Detection." *Psychological Review* 76 (3): 308. doi:10.1037/h0027238.

Kriegel, Uriah. 2009. *Subjective Consciousness: a Self-Representational Theory*. New York: Oxford University Press.

Kross, Ethan, Marc G Berman, Walter Mischel, Edward E Smith, and Tor D Wager. 2011. "Social Rejection Shares Somatosensory Representations with Physical Pain." *Proceedings of the National Academy of Sciences of the United States of America* 108 (15). National Academy of Sciences: 6270–75. doi:10.2307/41126642?ref=search-gateway:a3e9b807ff689aa6cb112a8d3300f51b.

Laeng, Bruno, and Unni Sulutvedt. 2014. "The Eye Pupil Adjusts to Imaginary Light." *Psychological Science* 25 (1): 188–97. doi:10.1177/0956797613503556.

Lak, Armin, Gil M Costa, Erin Romberg, Alexei A Koulakov, Zachary F Mainen, and Adam Kepecs. 2014. "Orbitofrontal Cortex Is Required for Optimal Waiting Based on Decision Confidence." *Neuron* 84: 1–12. doi:10.1016/j.neuron.2014.08.039.

Lamme, Victor A F. 2006. "Towards a True Neural Stance on Consciousness." *Trends in Cognitive Sciences* 10 (11): 494–501. doi:10.1016/j.tics.2006.09.001.

Lamme, Victor A F. 2010. "How Neuroscience Will Change Our View on Consciousness." *Cognitive Neuroscience* 1 (3): 204–20. doi:10.1080/17588921003731586.

Lamme, Victor A F, and Pieter R Roelfsema. 2000. "The Distinct Modes of Vision Offered by Feedforward and Recurrent Processing." *Trends in Neurosciences* 23 (11): 571–79. doi:10.1016/S0166-2236(00)01657-X.

Lamme, Victor A F, Karl Zipser, and Henk Spekreijse. 2002. "Masking Interrupts Figure-Ground Signals in V1." *Journal of Cognitive Neuroscience* 14 (7): 1044–53. doi:10.1162/089892902320474490.

Lamy, Dominique, Moti Salti, and Yair Bar-Haim. 2009. "Neural Correlates of Subjective Awareness and Unconscious Processing: an ERP Study." *Journal of Cognitive Neuroscience* 21 (7): 1435–46. doi:10.1162/jocn.2009.21064.

Landman, Rogier, Henk Spekreijse, and Victor A F Lamme. 2003. "Large Capacity Storage of Integrated Objects Before Change Blindness." *Vision Research* 43 (2): 149–64.

Langland-Hassan, Peter. 2017. "Pain and Incorrigibility." In *The Routledge Handbook of Philosophy of Pain*, edited by Jennifer Corns. London: Routledge.

Lau, Hakwan. 2008. "A Higher Order Bayesian Decision Theory of Consciousness." In *Progress in Brain Research*, edited by R Banerjee and B K Chakrabarti, 168:35–48. Progress in Brain Research. doi:10.1016/S0079-6123(07)68004-2.

Lau, Hakwan, and David Rosenthal. 2011. "Empirical Support for Higher-Order Theories of Conscious Awareness." *Trends in Cognitive Sciences* 15 (8): 365–73. doi:10.1016/j.tics.2011.05.009.

Lau, Hakwan, and R E Passingham. 2006. "Relative Blindsight in Normal Observers and the Neural Correlate of Visual Consciousness." *Proceedings of the National Academy of Sciences of the United States of America* 103 (49): 18763–68. doi:10.1073/pnas.0607716103.

Laureys, Steven, Gastone G Celesia, Francois Cohadon, Jan Lavrijsen, José León-Carrión, Walter G Sannita, Leon Sazbon, et al. 2010. "Unresponsive Wakefulness Syndrome: a New Name for the Vegetative State or Apallic Syndrome." *BMC Medicine* 8 (1): 68. doi:10.1186/1741-7015-8-68.

Lavie, N, D M Beck, and N Konstantinou. 2014. "Blinded by the Load: Attention, Awareness and the Role of Perceptual Load." *Philosophical Transactions of the Royal Society B: Biological Sciences* 369 (1641): 20130205–5. doi:10.1016/S0166-2236(96)20049-9.

Leeuw, Maaike, Mariëlle E J B Goossens, Steven J Linton, Geert Crombez, Katja Boersma, and Johan W S Vlaeyen. 2007. "The Fear-Avoidance Model of Musculoskeletal Pain: Current State of Scientific Evidence." *Journal of Behavioral Medicine* 30 (1): 77–94. doi:10.1007/s10865-006-9085-0.

Legrain, Valéry, Gian Domenico Iannetti, Léon Plaghki, and André Mouraux. 2011. "The Pain Matrix Reloaded: a Salience Detection System for the Body." *Progress in Neurobiology* 93 (1): 111–24. doi:10.1016/j.pneurobio.2010.10.005.

Legrain, Valéry, Raymond Bruyer, Jean-Michel Guérit, and Léon Plaghki. 2005. "Involuntary Orientation of Attention to Unattended Deviant Nociceptive Stimuli Is Modulated by Concomitant Visual Task Difficulty. Evidence From Laser Evoked Potentials." *Clinical Neurophysiology* 116 (9): 2165–74. doi:10.1016/j.clinph.2005.05.019.

Legrain, Valéry, Stefaan Van Damme, Christopher Eccleston, Karen D Davis, David A Seminowicz, and Geert Crombez. 2009. "A Neurocognitive Model of Attention to Pain: Behavioral and Neuroimaging Evidence." *Pain* 144 (3): 230–32. doi:10.1016/j.pain.2009.03.020.

Leopold, D A, and N K Logothetis. 1996. "Activity Changes in Early Visual Cortex Reflect Monkeys' Percepts During Binocular Rivalry." *Nature* 379 (6565): 549–53. doi:10.1038/379549a0.

Li, Q, Z Hill, and B J He. 2014. "Spatiotemporal Dissociation of Brain Activity Underlying Subjective Awareness, Objective Performance and Confidence." *The Journal of Neuroscience* 34 (12): 4382–95. doi:10.1523/JNEUROSCI.1820-13.2014.

Liu, Taosheng, Stuart Fuller, and Marisa Carrasco. 2006. "Attention Alters the Appearance of Motion Coherence." *Psychonomic Bulletin & Review* 13 (6): 1091–96.

Longuenesse, Béatrice. 1998. *Kant and the Capacity to Judge*. Translated by Charles T Wolfe. Princeton: Princeton University Press.

Luce, R Duncan. 1963. "Detection and Recognition." In *Handbook of Mathematical Psychology*, edited by R Duncan Luce, Robert R Bush, and Eugene Galanter, 103–89. New York: John Wiley & Sons.

Luce, R Duncan. 1990. "'On the Possible Psychophysical Laws' Revisited: Remarks on Cross-Modal Matching." *Psychological Review* 97 (1): 66–77.

Luck, S J, E K Vogel, and K L Shapiro. 1996. "Word Meanings Can Be Accessed but Not Reported During the Attentional Blink." *Nature* 383 (6601): 616–18. doi:10.1038/383616a0.

Lumer, E D, and G Rees. 1999. "Covariation of Activity in Visual and Prefrontal Cortex Associated with Subjective Visual Perception." *Proceedings of the National Academy of Sciences of the United States of America* 96 (4): 1669–73.

Machery, Edouard. 2014. "In Defense of Reverse Inference." *The British Journal for the Philosophy of Science* 65 (2): 251–67. doi:10.1093/bjps/axs044.

Macmillan, Neil A, and C Douglas Creelman. 2005. *Detection Theory*. 2nd ed. Mahwah, NJ: Lawrence Erlbaum.

Maniscalco, Brian, and Hakwan Lau. 2012. "A Signal Detection Theoretic Approach for Estimating Metacognitive Sensitivity From Confidence Ratings." *Consciousness and Cognition* 21: 422–30.

Maniscalco, Brian, and Hakwan Lau. 2016. "The Signal Processing Architecture Underlying Subjective Reports of Sensory Awareness." *Neuroscience of Consciousness* 2016 (1): 292. doi:10.1093/nc/niw002.

Mante, Valerio, David Sussillo, Krishna V Shenoy, and William T Newsome. 2013. "Context-Dependent Computation by Recurrent Dynamics in Prefrontal Cortex." *Nature* 503 (7474): 78–84. doi:10.1038/nature12742.

Marks, D F. 1973. "Visual Imagery Differences in the Recall of Pictures." *British Journal of Psychology* 64 (1): 17–24.

Martínez, Manolo. 2010. "Imperative Content and the Painfulness of Pain." *Phenomenology and the Cognitive Sciences* 10 (1): 67–90. doi:10.1007/s11097-010-9172-0.

McGinn, Colin. 2004. *Mindsight: Image, Dream, Meaning*. Cambridge, MA: Harvard University Press.

Melnick, Michael D, Duje Tadin, and Krystel R Huxlin. 2016. "Relearning to See in Cortical Blindness." *The Neuroscientist: A Review Journal Bringing Neurobiology, Neurology and Psychiatry* 22 (2): 199–212. doi:10.1177/1073858415621035.

Melzack, R. 1975. "The McGill Pain Questionnaire: Major Properties and Scoring Methods." *Pain* 1 (3): 277–99.

Merikle, P M, D Smilek, and J D Eastwood. 2001. "Perception Without Awareness: Perspectives From Cognitive Psychology." *Cognition* 79 (1-2): 115–34.

Mettler, Frederick Albert. 1949. *Selective Partial Ablation of the Frontal Cortex*. Hoeber.

Michel, Matthias. 2017. "A Role for the Anterior Insular Cortex in the Global Neuronal Workspace Model of Consciousness." *Consciousness and Cognition* 49 (March): 333–46. doi:10.1016/j.concog.2017.02.004.

Miller, E K. 2000. "The Prefrontal Cortex and Cognitive Control." *Nature Reviews Neuroscience* 1 (1): 59–65. doi:10.1038/35036228.

Miller, E K, and J D Cohen. 2001. "An Integrative Theory of Prefrontal Cortex Function." *Annual Review of Neuroscience* 24: 167–202. doi:10.1146/annurev.neuro.24.1.167.

Milner, A D A. David, and Melvyn A Goodale. 2006. *The Visual Brain in Action*. 2nd ed. Oxford ; New York : Oxford University Press.

Miron, D, G H Duncan, and M C Bushnell. 1989. "Effects of Attention on the Intensity and Unpleasantness of Thermal Pain." *Pain* 39 (3): 345–52.

Miyamoto, Kentaro, Takahiro Osada, Rieko Setsuie, Masaki Takeda, Keita Tamura, Yusuke Adachi, and Yasushi Miyashita. 2017. "Causal Neural Network of Metamemory for Retrospection in Primates." *Science* 355 (6321): 188–93. doi:10.1126/science.aal0162.

Moayedi, M, and K D Davis. 2012. “Theories of Pain: From Specificity to Gate Control.” *Journal of Neurophysiology* 109 (1): 5–12. doi:10.1152/jn.00457.2012.

Montagna, Barbara, and Marisa Carrasco. 2006. “Transient Covert Attention and the Perceived Rate of Flicker.” *Journal of Vision* 6 (9): 955–65. doi:10.1167/6.9.8.

Montague, Michelle. 2015. “Cognitive Phenomenology and Conscious Thought.” *Phenomenology and the Cognitive Sciences* 15 (2): 167–81. doi:10.1111/phc3.12053.

Montemayor, Carlos, and Harry Haroutioun Haladjian. 2015. *Consciousness, Attention, and Conscious Attention*. Cambridge, MA: MIT Press.

Morales, Jorge, Guillermo Solovey, Brian Maniscalco, Dobromir A Rahnev, Floris P de Lange, and Hakwan Lau. 2015. “Low Attention Impairs Optimal Incorporation of Prior Knowledge in Perceptual Decisions.” *Attention, Perception, & Psychophysics* 77 (6): 2021–36. doi:10.3758/s13414-015-0897-2.

Morales, Jorge, Hakwan Lau, and Stephen M Fleming. 2018. “Domain-General and Domain-Specific Patterns of Activity Supporting Metacognition in Human Prefrontal Cortex.” *The Journal of Neuroscience*. doi:10.1101/172445.

Morrison, John. 2016. “Perceptual Confidence.” *Analytic Philosophy* 57 (1): 15–48.

Morrison, John. 2017. “Perceptual Confidence and Categorization.” *Analytic Philosophy*, January, 1–12.

Mulder, Martijn J, Eric-Jan Wagenmakers, Roger Ratcliff, Wouter Boekel, and Birte U Forstmann. 2012. “Bias in the Brain: a Diffusion Model Analysis of Prior Probability and Potential Payoff.” *The Journal of Neuroscience* 32 (7): 2335–43. doi:10.1523/JNEUROSCI.4156-11.2012.

Neisser, Joseph. 2012. “Consciousness and Cognition.” *Consciousness and Cognition* 21 (2): 681–90. doi:10.1016/j.concog.2011.03.012.

Newman, Lex. 2016. “Descartes’ Epistemology.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N Zalta. https://plato.stanford.edu/archives/win2016/entries/descartes-epistemology/.

Nikolajsen, L, and T S Jensen. 2001. “Phantom Limb Pain.” *British Journal of Anaesthesia* 87 (1): 107–16.

Noë, Alva, and E Thompson. 2004. “Are There Neural Correlates of Consciousness?” *Journal of Consciousness Studies* 11 (1): 3–28.

Noy, N, S Bickel, E Zion-Golumbic, M Harel, T Golan, I Davidesco, C A Schevon, et al. 2015. “Ignition’s Glow: Ultra-Fast Spread of Global Cortical Activity Accompanying Local “Ignitions” in Visual Cortex During Conscious Visual Perception.” *Consciousness and Cognition* 35 (September): 206–24. doi:10.1016/j.concog.2015.03.006.

Odegaard, Brian, Robert T Knight, and Hakwan Lau. 2017. “Should a Few Null Findings Falsify Prefrontal Theories of Conscious Perception?” *The Journal of Neuroscience* 37 (40): 9593–9602. doi:10.1101/122267.

Overgaard, M, and K Sandberg. 2012. “Kinds of Access: Different Methods for Report Reveal Different Kinds of Metacognitive Access.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 367 (1594): 1287–96. doi:10.1038/nn0207-140.

Owen, A M. 2006. “Detecting Awareness in the Vegetative State.” *Science* 313 (5792): 1402–2. doi:10.1126/science.1130197.

Panagiotaropoulos, Theofanis I, Gustavo Deco, Vishal Kapoor, and Nikos K Logothetis. 2012. “Neuronal Discharges and Gamma Oscillations Explicitly Reflect Visual Consciousness in the Lateral Prefrontal Cortex.” *Neuron* 74 (5): 924–35. doi:10.1016/j.neuron.2012.04.013.

Passingham, Richard. 2009. “How Good Is the Macaque Monkey Model of the Human Brain?” *Genetic and Optical Targeting of Neural Circuits and Behavior —Zebrafish in the Spotlight* 19 (1): 6–11. doi:10.1016/j.conb.2009.01.002.

Passingham, Richard E, and Steven P Wise. 2012. *The Neurobiology of the Prefrontal Cortex*. Oxford: Oxford University Press.

Peacocke, Christopher. 1992. “Scenarios, Concepts and Perception.” In *The Contents of Experience*, edited by Tim Crane, 105–35. Cambridge: Cambridge University Press.

Peacocke, Christopher. 1998. “Conscious Attitudes, Attention and Self-Knowledge.” In *Knowing Our Own Minds*, edited by Crispin Wright, Barry C Smith, and Cynthia Macdonald, 63–98. New York: Oxford University Press.

Peacocke, Christopher. 2015. “Magnitudes: Metaphysics, Explanation, and Perception.” In *Mind, Language, Action: Proceedings of the 2013 Kirchberg Symposium*, edited by D Moyal-Sharrock and V Munz, 357–87. Berlin: de Gruyter.

Pearson, J, R L Rademaker, and F Tong. 2011. “Evaluating the Mind’s Eye: the Metacognition of Visual Imagery.” *Psychological Science* 22 (12): 1535–42. doi:10.1177/0956797611417134.

Pearson, Joel, Thomas Naselaris, Emily A Holmes, and Stephen M Kosslyn. 2015. “Mental Imagery: Functional Mechanisms and Clinical Applications.” *Trends in Cognitive Sciences* 19 (10): 590–602. doi:10.1016/j.tics.2015.08.003.

Persaud, Navindra, Matthew Davidson, Brian Maniscalco, Dean Mobbs, Richard E Passingham, Alan Cowey, and Hakwan Lau. 2011. “Awareness-Related Activity in Prefrontal and Parietal Cortices in Blindsight Reflects More Than Superior Visual Performance.” *NeuroImage* 58 (2): 605–11. doi:10.1016/j.neuroimage.2011.06.081.

Pessoa, Luiz, and Ralph Adolphs. 2010. “Emotion Processing and the Amygdala: From a ‘Low Road’ to ‘Many Roads’ of Evaluating Biological Significance.” *Nature Reviews Neuroscience* 11 (11): 773–83. doi:10.1038/nrn2920.

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.” *Neuron* 72 (5): 832–46. doi:10.1016/j.neuron.2011.09.025.

Peters, Megan A K, and Hakwan Lau. 2015. “Human Observers Have Optimal Introspective Access to Perceptual Processes Even for Visually Masked Stimuli.” *eLife* 4. doi:10.7554/eLife.09651.

Petrides, M, and D N Pandya. 1984. “Projections to the Frontal Cortex From the Posterior Parietal Region in the Rhesus Monkey.” *The Journal of Comparative Neurology* 228 (1): 105–16. doi:10.1002/cne.902280110.

Phillips, Ian. 2016. “Consciousness and Criterion: on Block’s Case for Unconscious Seeing.” *Philosophy and Phenomenological Research* 93 (2): 419–51. doi:10.1111/phpr.12224.

Picciuto, Vincent J, and Peter Carruthers. 2014. “Inner Sense.” In *Perception and Its Modalities*, edited by Dustin Stokes, Mohan Matthen, and Stephen Biggs, 277–96. New York: Oxford University Press.

Pins, D, and D ffytche. 2003. “The Neural Correlates of Conscious Vision.” *Cerebral Cortex* 13 (5): 461–74.

Pitts, Michael A, Jennifer Padwal, Daniel Fennelly, Antígona Martínez, and Steven A Hillyard. 2014. “Gamma Band Activity and the P3 Reflect Post-Perceptual Processes, Not Visual Awareness.” *NeuroImage* 101 (November): 337–50. doi:10.1016/j.neuroimage.2014.07.024.

Pleskac, Timothy J, and Jerome R Busemeyer. 2010. "Two-Stage Dynamic Signal Detection: a Theory of Choice, Decision Time, and Confidence." *Psychological Review* 117 (3): 864–901. doi:10.1037/a0019737.

Poldrack, Russell A. 2006. "Can Cognitive Processes Be Inferred From Neuroimaging Data?" *Trends in Cognitive Sciences* 10 (2): 59–63. doi:10.1016/j.tics.2005.12.004.

Posner, M I, C R Snyder, and B J Davidson. 1980. "Attention and the Detection of Signals." *Journal of Experimental Psychology: General* 109 (2): 160–74.

Posner, Michael I. 1980. "Orienting of Attention." *Quarterly Journal of Experimental Psychology* 32 (1): 3–25. doi:10.1080/00335558008248231.

Prescott, Steven A, Qiufu Ma, and Yves De Koninck. 2014. "Normal and Abnormal Coding of Somatosensory Stimuli Causing Pain." *Nature Neuroscience* 17 (2): 183–91. doi:10.1038/nn.3629.

Prinz, Jesse. 2005. "Are Emotions Feelings?" *Journal of Consciousness Studies* 12 (8-10): 9–25.

Prinz, Jesse. 2012. *The Conscious Brain*. New York: Oxford University Press.

Rahnev, Dobromir A, Brian Maniscalco, Tashina Graves, Elliott Huang, Floris P de Lange, and Hakwan Lau. 2011. "Attention Induces Conservative Subjective Biases in Visual Perception." *Nature Reviews: Neuroscience* 14 (12): 1513–15. doi:10.1038/nn.2948.

Railo, Henry, Mika Koivisto, and Antti Revonsuo. 2011. "Tracking the Processes Behind Conscious Perception: a Review of Event-Related Potential Correlates of Visual Consciousness." *Consciousness and Cognition* 20 (3): 972–83. doi:10.1016/j.concog.2011.03.019.

Rainville, P. 2002. "Brain Mechanisms of Pain Affect and Pain Modulation." *Current Opinion in Neurobiology* 12 (2): 195–204.

Rainville, P, B Carrier, R K Hofbauer, M C Bushnell, and G H Duncan. 1999. "Dissociation of Sensory and Affective Dimensions of Pain Using Hypnotic Modulation." *Pain* 82 (2): 159–71.

Reuter, K. 2011. "Distinguishing the Appearance From the Reality of Pain." *Journal of Consciousness Studies*.

Reynolds, John H, and David J Heeger. 2009. "The Normalization Model of Attention." *Neuron* 61 (2): 168–85. doi:10.1016/j.neuron.2009.01.002.The.

Rigotti, Mattia, Omri Barak, Melissa R Warden, Xiao-Jing Wang, Nathaniel D Daw, Earl K Miller, and Stefano Fusi. 2013. "The Importance of Mixed Selectivity in Complex Cognitive Tasks." *Nature* 497 (7451): 585–90. doi:10.1038/nature12160.

Robinson, Z, C J Maley, and Gualtiero Piccinini. 2015. "Is Consciousness a Spandrel?" *Journal of the American Philosophical Association* 1 (2): 365–83. doi:10.1017/apa.2014.10.

Romanski, L M, M Giguere, J F Bates, and P S Goldman-Rakic. 1997. "Topographic Organization of Medial Pulvinar Connections with the Prefrontal Cortex in the Rhesus Monkey." *The Journal of Comparative Neurology* 379 (3): 313–32.

Rorty, Richard. 1970. "Incorrigibility as the Mark of the Mental." *The Journal of Philosophy* 67 (12). Journal of Philosophy: 399–424.

Rosenthal, D M. 1993. "State Consciousness and Transitive Consciousness." *Consciousness and Cognition* 2: 355–63.

Rosenthal, David. 2005. *Consciousness and Mind*. New York: Oxford University Press.

Rosenthal, David. 2008. "Consciousness and Its Function." *Neuropsychologia* 46 (3): 829–40. doi:10.1016/j.neuropsychologia.2007.11.012.

Rosenthal, David. 2018. "Consciousness and Confidence." *Neuropsychologia*, February. doi:10.1016/j.neuropsychologia.2018.01.018.

Rounis, Elisabeth, Brian Maniscalco, John C Rothwell, Richard E Passingham, and Hakwan Lau. 2010. "Theta-Burst Transcranial Magnetic Stimulation to the Prefrontal Cortex Impairs Metacognitive Visual Awareness." *Cognitive Neuroscience* 1 (3): 165–75. doi:10.1080/17588921003632529.

Rutiku, R, M Martin, T Bachmann, and J Aru. 2015. "Does the P300 Reflect Conscious Perception or Its Consequences?" *Neuroscience* 298: 180–89. doi:10.1016/j.neuroscience.2015.04.029.

Ryle, Gilbert. 2009. *The Concept of Mind*. New York: Routledge.

Salti, Moti, Yair Bar-Haim, and Dominique Lamy. 2012. "The P3 Component of the ERP Reflects Conscious Perception, Not Confidence." *Consciousness and Cognition* 21 (2): 961–68. doi:10.1016/j.concog.2012.01.012.

Samaha, Jason, John J Barrett, Andrew D Sheldon, Joshua J LaRocque, and Bradley R Postle. 2016. "Dissociating Perceptual Confidence From Discrimination Accuracy Reveals No Influence of Metacognitive Awareness on Working Memory." *Frontiers in Psychology* 7 (938): 166. doi:10.3389/fnint.2012.00079.

Sandberg, Kristian, Bert Timmermans, Morten Overgaard, and Axel Cleeremans. 2010. "Measuring Consciousness: Is One Measure Better Than the Other?" *Consciousness and Cognition* 19 (4): 1069–78. doi:10.1016/j.concog.2009.12.013.

Sandberg, Kristian, Bo Martin Bibby, Bert Timmermans, Axel Cleeremans, and Morten Overgaard. 2011. "Measuring Consciousness: Task Accuracy and Awareness as Sigmoid Functions of Stimulus Duration." *Consciousness and Cognition* 20 (4): 1659–75. doi:10.1016/j.concog.2011.09.002.

Sandrini, Marco, Carlo Umiltá, and Elena Rusconi. 2011. "The Use of Transcranial Magnetic Stimulation in Cognitive Neuroscience: a New Synthesis of Methodological Issues."

Neuroscience and Biobehavioral Reviews 35 (3): 516–36. doi:10.1016/j.neubiorev.2010.06.005.

Sartre, Jean-Paul. 2004. *The Imaginary*. Translated by Jonathan Webber. New York: Routledge.

Savage, C Wade. 1970. *The Measurement of Sensation*. Berkeley: University of California Press.

Schmid, Michael C, Sylwia W Mrowka, Janita Turchi, Richard C Saunders, Melanie Wilke, Andrew J Peters, Frank Q Ye, and David A Leopold. 2010. "Blindsight Depends on the Lateral Geniculate Nucleus." *Nature* 466 (7304): 373–77. doi:10.1038/nature09179.

Schooler, Jonathan W, Jonathan Smallwood, Kalina Christoff, Todd C Handy, Erik D Reichle, and Michael A Sayette. 2011. "Meta-Awareness, Perceptual Decoupling and the Wandering Mind." *Trends in Cognitive Sciences* 15 (7): 319–26. doi:10.1016/j.tics.2011.05.006.

Schwitzgebel, Eric. 2008. "The Unreliability of Naive Introspection." *Philosophical Review* 117 (2). Duke University Press: 245–73.

Schwitzgebel, Eric. 2011. *Perplexities of Consciousness*. Cambridge, MA: MIT Press.

Schwitzgebel, Eric. 2012. "Introspection, What?" In *Introspection and Consciousness*, edited by Declan Smithies and Daniel Stoljar, 29–48. New York: Oxford University Press.

Sergent, Claire, Sylvain Baillet, and Stanislas Dehaene. 2005. "Timing of the Brain Events Underlying Access to Consciousness During the Attentional Blink." *Nature Neuroscience* 8 (10): 1391–1400. doi:10.1038/nn1549.

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." *Current Biology* 23 (2): 150–55. doi:10.1016/j.cub.2012.11.047.

Seth, Anil K, Zoltán Dienes, Axel Cleeremans, Morten Overgaard, and Luiz Pessoa. 2008. "Measuring Consciousness: Relating Behavioural and Neurophysiological Approaches." *Trends in Cognitive Sciences* 12 (8): 314–21. doi:10.1016/j.tics.2008.04.008.

Shadlen, Michael N, and Roozbeh Kiani. 2013. "Decision Making as a Window on Cognition." *Neuron* 80 (3): 791–806. doi:10.1016/j.neuron.2013.10.047.

Shipp, S. 2003. "The Functional Logic of Cortico-Pulvinar Connections." *Philosophical Transactions of the Royal Society B: Biological Sciences* 358 (1438). The Royal Society: 1605–24. doi:10.1098/rstb.2002.1213.

Shoemaker, Sydney S. 1996. *The First-Person Perspective and Other Essays*. New York: Cambridge University Press.

Shoemaker, Sydney S. 2000. "Introspection and Phenomenal Character." *Philosophical Topics* 28 (2): 247–73.

Shoemaker, Sydney S. 1981. "Some Varieties of Functionalism." *Philosophical Topics* 12 (1). Cengage Learning: 93–119.

Sligte, Ilja G, H Steven Scholte, and Victor A F Lamme. 2008. "Are There Multiple Visual Short-Term Memory Stores?" Edited by Sheng He. *PLoS ONE* 3 (2): e1699. doi:10.1371/journal.pone.0001699.g008.

Smith, J David. 2009. "The Study of Animal Metacognition." *Trends in Cognitive Sciences* 13 (9): 389–96. doi:10.1016/j.tics.2009.06.009.

Snodgrass, Michael, and Howard Shevrin. 2006. "Unconscious Inhibition and Facilitation at the Objective Detection Threshold: Replicable and Qualitatively Different Unconscious Perceptual Effects." *Cognition* 101 (1): 43–79. doi:10.1016/j.cognition.2005.06.006.

Snyder, Solomon H. 1996. *Drugs and the Brain*. New York: W. H. Freeman & Co.

Spener, Maja. 2015. “Calibrating Introspection.” *Philosophical Issues* 25 (1): 300–321. doi:10.1111/phis.12062.

Stazicker, James. 2011. “Attention, Visual Consciousness and Indeterminacy.” *Mind & Language* 26 (2): 156–84.

Stokes, Mark G. 2015. “‘Activity-Silent’ Working Memory in Prefrontal Cortex: a Dynamic Coding Framework.” *Trends in Cognitive Sciences* 19 (7): 394–405. doi:10.1016/j.tics.2015.05.004.

Stroud, Barry. 1977. *Hume*. London: Routledge and Kegan Paul.

Sullivan, M J, B Thorn, J A Haythornthwaite, F Keefe, M Martin, L A Bradley, and J C Lefebvre. 2001. “Theoretical Perspectives on the Relation Between Catastrophizing and Pain.” *The Clinical Journal of Pain* 17 (1): 52–64.

Sun, J, and P Perona. 1998. “Where Is the Sun?” *Nature Neuroscience* 1 (3): 183–84. doi:10.1038/630.

Swets, J A. 1961. “Is There a Sensory Threshold?” *Science* 134 (3473): 168–77.

Thomas, Nigel. 2009. “Visual Imagery and Consciousness.” In *Encyclopedia of Consciousness*, edited by W P Banks, 445–57. Academic Press/Elsevier.

Tononi, Giulio. 2008. “Consciousness as Integrated Information.” *Biology Bulletin* 215 (December): 216–42.

Tononi, Giulio, Melanie Boly, Marcello Massimini, and Christof Koch. 2016. “Integrated Information Theory: From Consciousness to Its Physical Substrate.” *Nature Reviews Neuroscience* 17 (7): 450–61. doi:10.1038/nrn.2016.44.

Tse, P U, S Martinez-Conde, A A Schlegel, and S L Macknik. 2005. “Visibility, Visual Awareness, and Visual Masking of Simple Unattended Targets Are Confined to Areas in the Occipital Cortex Beyond Human V1/V2.” *Proceedings of the National Academy*

of Sciences of the United States of America 102 (47): 17178–83. doi:10.1073/pnas.0508010102.

Tse, Peter U. 2005. "Voluntary Attention Modulates the Brightness of Overlapping Transparent Surfaces." *Vision Research* 45 (9): 1095–98. doi:10.1016/j.visres.2004.11.001.

Tsuchiya, Naotsugu, and Christof Koch. 2005. "Continuous Flash Suppression Reduces Negative Afterimages." *Nature Neuroscience* 8 (8): 1096–1101. doi:10.1038/nn1500.

Tsuchiya, Naotsugu, Melanie Wilke, Stefan Frässle, and Victor A F Lamme. 2015. "No-Report Paradigms: Extracting the True Neural Correlates of Consciousness." *Trends in Cognitive Sciences* 19 (12): 757–70. doi:10.1016/j.tics.2015.10.002.

Tye, Michael. 1992. "Visual Qualia and Visual Content." In *The Contents of Experience*, edited by Tim Crane, 158–76. Cambridge: Cambridge University Press.

Tye, Michael. 1995. "A Representational Theory of Pains and Their Phenomenal Character." *Philosophical Perspectives* 9: 223–39. doi:10.2307/2214219.

Tye, Michael. 1996. "The Function of Consciousness." *Noûs* 30 (3): 287–305.

Tye, Michael. 2016. "Qualia." Edited by Edward N Zalta. *Stanford Encyclopedia of Philosophy*, Winter. https://plato.stanford.edu/archives/win2016/entries/qualia/.

Tye, Michael. 2000. *Consciousness, Color, and Content*. Cambridge, MA: MIT Press.

Tye, Michael. 2002. "Representationalism and the Transparency of Experience." *Noûs* 36 (1). Wiley: 137–51. doi:10.2307/3506107.

Ungerleider, Leslie G, Mortimer Mishkin, and R J W Mansfield. 1982. "Two Cortical Visual Systems." In *Analysis of Visual Behavior*, edited by D J Ingle and Melvyn A Goodale, 548–86.

Van Opstal, Filip, Floris P de Lange, and Stanislas Dehaene. 2011. "Rapid Parallel Semantic Processing of Numbers Without Awareness." *Cognition* 120 (1): 136–47. doi:10.1016/j.cognition.2011.03.005.

van Ravenzwaaij, Don, Martijn J Mulder, Francis Tuerlinckx, and Eric-Jan Wagenmakers. 2012. "Do the Dynamics of Prior Information Depend on Task Context? an Analysis of Optimal Performance and an Empirical Test." *Frontiers in Psychology* 3 (May). Frontiers Media SA. doi:10.3389/fpsyg.2012.00132.

Vandenbroucke, Annelinde R E, Ilja G Sligte, Jade G de Vries, Michael X Cohen, and Victor A F Lamme. 2015. "Neural Correlates of Visual Short-Term Memory Dissociate Between Fragile and Working Memory Representations." *J. Cognitive Neuroscience* 27 (12): 2477–90. doi:10.1162/jocn_a_00870.

Vlassova, A, C Donkin, and J Pearson. 2014. "Unconscious Information Changes Decision Accuracy but Not Confidence." *Proceedings of the National Academy of Sciences of the United States of America*, October. doi:10.1073/pnas.1403619111.

Wager, Tor D, Lauren Y Atlas, Martin A Lindquist, Mathieu Roy, Choong-Wan Woo, and Ethan Kross. 2013. "An fMRI-Based Neurologic Signature of Physical Pain." *New England Journal of Medicine* 368 (15): 1388–97. doi:10.1056/NEJMoa1204471.

Wald, Abraham. 1947. *Sequential Analysis*. New York: John Wiley & Sons.

Wales, R, and R Fox. 1970. "Increment Detection Thresholds During Binocular Rivalry Suppression." *Perception & Psychophysics* 8 (2): 90–94.

Wang, Megan, Daniel Arteaga, and Biyu J He. 2013. "Brain Mechanisms for Simple Perception and Bistable Perception" 110 (35): E3350–59. doi:10.1073/pnas.1221945110/-/DCSupplemental.

Watzl, Sebastian. 2017. *Structuring Mind*. Oxford: Oxford University Press.

Weiskrantz, L. 1986. *Blindsight: a Case Study and Implications*. Oxford: Oxford University Press.

Weiskrantz, L, J L Barbur, and A Sahraie. 1995. "Parameters Affecting Conscious Versus Unconscious Visual Discrimination with Damage to the Visual Cortex (V1)." *Proceedings of the National Academy of Sciences of the United States of America* 92 (13): 6122–26.

Witt, Jessica K, J Eric T Taylor, Mila Sugovic, and John T Wixted. 2015. "Signal Detection Measures Cannot Distinguish Perceptual Biases From Response Biases." *Perception* 44 (3): 289–300. doi:10.1068/p7908.

Wixted, John T. 2004. "In Defense of the Signal Detection Interpretation of Remember / Know Judgments." *Psychonomic Bulletin & Review* 11 (4): 616–41.

Wixted, John T. 2009. "Remember/Know Judgments in Cognitive Neuroscience: an Illustration of the Underrepresented Point of View." *Learning & Memory (Cold Spring Harbor, N.Y.)* 16 (7): 406–12. doi:10.1101/lm.1312809.

Wixted, John T, and Laura Mickes. 2010. "A Continuous Dual-Process Model of Remember/Know Judgments." *Psychological Review* 117 (4): 1025–54. doi:10.1037/a0020874.

Woo, Choong-Wan, Leonie Koban, Ethan Kross, Martin A Lindquist, Marie T Banich, Luka Ruzic, Jessica R Andrews-Hanna, and Tor D Wager. 2014. "Separate Neural Representations for Physical Pain and Social Rejection." *Nature Communications* 5 (November): 1–12. doi:10.1038/ncomms6380.

Wright, Richard D, and Lawrence M Ward. 2008. *Orienting of Attention*. New York: Oxford University Press.

Wu, Wayne. 2011. "What Is Conscious Attention?" *Philosophy and Phenomenological Research* 82 (1): 93–120.

Wu, Wayne. 2014a. “Against Division: Consciousness, Information and the Visual Streams.” *Mind & Language*.

Wu, Wayne. 2014b. *Attention*. New York: Routledge.

Wu, Wayne. 2017. “Shaking Up the Mind’s Ground Floor: The Cognitive Penetration of Visual Attention.” *The Journal of Philosophy* 114 (1): 5–32. doi:10.5840/jphil201711411.

Yonelinas, Andrew P, and Larry L Jacoby. 2012. “The Process-Dissociation Approach Two Decades Later: Convergence, Boundary Conditions, and New Directions.” *Memory & Cognition* 40 (5): 663–80. doi:10.3758/s13421-012-0205-5.

Zeidan, Fadel, Nakia S Gordon, Junaid Merchant, and Paula Goolkasian. 2010. “The Effects of Brief Mindfulness Meditation Training on Experimentally Induced Pain.” *The Journal of Pain : Official Journal of the American Pain Society* 11 (3): 199–209. doi:10.1016/j.jpain.2009.07.015.

Zeidan, Fadel, Nichole M Emerson, Suzan R Farris, Jenna N Ray, Youngkyoo Jung, John G McHaffie, and Robert C Coghill. 2015. “Mindfulness Meditation-Based Pain Relief Employs Different Neural Mechanisms Than Placebo and Sham Mindfulness Meditation-Induced Analgesia.” *The Journal of Neuroscience* 35 (46). Society for Neuroscience: 15307–25. doi:10.1523/JNEUROSCI.2542-15.2015.

Zeman, Adam, Michaela Dewar, and Sergio Della Sala. 2015. “Lives Without Imagery - Congenital Aphantasia.” *Cortex* 73 (December): 378–80. doi:10.1016/j.cortex.2015.05.019.

Appendix

Here I describe the technical details of the 2-choice discrimination simulation reported in Figures 8-11. All simulations were made in Matlab (MathWorks, Natick, MA). Equal Gaussian distributions were assumed for the internal response. The discrimination criterion (vertical solid line, Fig. 6) was simply the intersection of the two curves, which were arbitrarily placed at zero. The distributions for stimuli A and B were centered at $\mu=\pm d'/2$, respectively, where sensitivity measure d' equals 1 in Figures 8-10, it equals 2 in Figure 11a-b and it equals 0.5 in Figure 11c-d. The awareness criteria were arbitrarily set at ±2 (vertical dashed lines, Fig. 6) for Figures 8-10 and to ±3 in Figure 11a-b and ±0.5 in Figure 11c-d. The activation waveform was stipulated to be a simple sine wave from 0 to 2π (Fig. 8a-c). For simplicity, the domain was extended to 3π, in which an additional wave was added to represent the extra activity in the *aware* condition (Fig. 8a). The domain was scaled to 500ms to maintain consistency with ERPs. The sine wave amplitude was directly proportional to the internal response. Specifically, the goal of the analysis was to recover the response that only appears from 2π to 3π (i.e., 333 ms to 500 ms), which was stipulated to be specific to awareness. Below, in Equation 1, $A_{(unaware)}$ and $A_{(aware)}$ are the amplitudes of the waveforms associated with unaware and aware trials at every time point, respectively. Finally, x represents the internal perceptual response, and t represents time in milliseconds.

$$A_{(unaware)} = \begin{cases} x \sin t, & 0 < t < 333 \\ 0, & t > 333 \end{cases} \quad A_{(aware)} = x \sin t \qquad \text{Eq. (1)}$$

10,000 trials were used per simulation (Fig. 8) and performed LSB's correction method as presented in their endnote 2. The results of this correction are presented in Figures 9 and 11, whereas the results of the SDT-based correction method appear in Figures 10 and 11. They derived the estimated waveform corresponding to unaware-correct chance-free trials as follows:

$$A(UC_{chance-free}) = \frac{A(UC_{observed}) - \%UC_{chance} * A(UI_{observed})}{1 - \%UC_{chance}} \quad \text{Eq. (2)}$$

where A is the amplitude of the waveform at each time point, $UC_{observed}$ are the unconscious-correct trials, $UI_{observed}$ are the unconscious incorrect trials, and $\%UC_{chance}$ is the expected percentage of correct trials by chance during the unaware condition such that:

$$\%UC_{chance} = \frac{\%Unaware\ Incorrect\ trials}{\%\ Expected\ Incorrect\ trials\ by\ chance} * \%\ Expected\ correct\ trials\ by\ chance \quad \text{Eq. (3)}$$

Since we simulated, for computational simplicity, a 2-choice discrimination task rather than a 4-AFC task as was done in LSB, chance performance was 50% in all our simulations rather than 25%.

In the proposed correction method the awareness criteria can be inferred using standard Signal Detection Theory (SDT) (Maniscalco and Lau 2012). Note that the awareness criteria are slightly different from the standard discrimination criterion, but they are criteria all the same. So, in Equation 2 below the awareness criterion *ac* is determined by means of *C* and *I*, which represent the proportion of aware correct responses (analogous to hit rate) and the proportion of aware incorrect responses (analogous to false alarm rate), respectively. By using

these rates instead of the standard hit rate versus false alarm rate calculation, awareness criteria (*ac*) were determined in the same way that the discrimination criterion is determined. Thus,

$$ac = -\frac{z(C) + z(I)}{2} \qquad \text{Eq. (4)}$$

where z(*C*) and z(*I*) are the z-scores of *C* and *I*.

Because we worked in standardized space (i.e., the standard deviation of the Gaussians is 1), knowing *ac* allows us to estimate the mean internal perceptual response for each partition (Fig. 6) using expected value:

$$E(X) = \int x * p(x) = \frac{\int x * f(x)}{\int f(x)}\, dx \qquad \text{Eq. (5)}$$

This leaves us with the mean internal response for Stimulus A aware, Stimulus A unaware, Stimulus B aware, and Stimulus B unaware. Below, in Equation 6, 0 is the discrimination criterion, *ac*, as defined above, represents the awareness criterion, and *x* represents the internal perceptual response. Because the area under the curve of each partition is not 1, the means must be normalized. The equations are illustrated with Stimulus A as in Figure 2 (but it can trivially be altered for Stimulus B), and assumed *f(x)* to be a standard Gaussian distribution.

$$M_{unaware} = \frac{\int_0^{ac} x * f(x|stimulus = A) dx}{\int_0^{ac} f(x|stimulus = A)\, dx} \, M_{aware} = \frac{\int_{ac}^{-\infty} x * f(x|stimulus = A) dx}{\int_{ac}^{-\infty} f(x|stimulus = A) dx} \qquad \text{Eq. (6)}$$

The ratio of the means is used as a way to scale the waveform associated with the awareness condition and compare it to the unawareness condition to adjust the unaware waveform:

$$A_{(unaware\ SDT-adjusted)} = \frac{M_{aware}}{M_{unaware}} * A_{(unaware)} \qquad \text{Eq. (7)}$$

Finally, the $A_{(unaware\ SDT\text{-}adjusted)}$ waveform was subtracted from the $A_{(aware)}$ waveform to obtain the distinctive awareness waveform thus eliminating the potential performance confound.