PRIOR EXPECTATIONS SHAPE SUBJECTIVE PERCEPTUAL CONFIDENCE

Maxine T. Sherman

Thesis submitted for the degree of

Doctor of Philosophy

School of Psychology

University of Sussex

May 2016

DECLARATION

I hereby declare that this thesis has not been and will not be, submitted in whole or in part to another University for the award of any other degree.

Maxine T. Sherman

06/05/2016

PUBLICATIONS

The thesis conforms to an ‘article format’ thesis, in which the middle chapters consist of discrete articles written in a style that is appropriate for publication in peer-reviewed journals in the field. The first, second and final chapters present synthetic overviews and discussions of the field and the research undertaken.

The following chapters from this thesis have been published or are under review:

Chapter 3

This has been adapted from the following published book chapter:

Sherman, M. T., Barrett, A. B., & Kanai, R. (2015). Inferences about consciousness using subjective reports of confidence. In M. Overgaard (Ed.), *Behavioral Methods in Consciousness Research* (pp. 87–106). Oxford University Press

Author contributions: MS wrote the chapter; ABB and RK provided critical revisions.

Chapter 4

This has been published as:

Sherman, M. T., Seth, A. K., Barrett, A. B. & Kanai, R. (2015) Prior expectations facilitate metacognition for perceptual decision. *Consciousness and cognition,* 35, 53-65

Author contributions: MS conceived the study and was responsible for all data collection, all data analysis and writing the manuscript; RK provided feedback on the study design; ABB was responsible for modelling the data and simulating the predicted results; MS wrote the paper; AKS, RK and ABB provided critical revisions to the manuscript.

Chapter 5

This is published as:

Sherman, M.T**.**, Kanai, R., Seth, A.K., VanRullen, R. (2016) Rhythmic influence of top-down perceptual priors in the phase of pre-stimulus occipital alpha oscillations. *Journal of Cognitive Neuroscience,* 29, *1318-1330.*

Author contributions: MS conceived the study and was responsible for all data collection, all data analysis and writing the manuscript; RV, AS and RK provided feedback on the study design, guidance on EEG data analysis and critical revisions to the manuscript.

Chapter 6

This is in press as:

Sherman, M.T**.**, Seth A.K., Kanai, R. (in press). Predictions shape decision confidence in right inferior frontal gyrus. *Journal of Neuroscience.*

This paper has been uploaded as a pre-print onto BioRxiv, dx.doi.org/10.1101/047126

Author contributions: MS conceived the study and was responsible for all data collection, all data analysis and writing the manuscript; RK provided feedback on the study design; RK and AKS provided feedback on fMRI data analysis and provided critical revisions to the manuscript.

For my wonderful sister Alix,
I love you always.

"Intuition and concepts constitute... the elements of all our knowledge, so that neither concepts without an intuition in some way corresponding to them, nor intuition without concepts, can yield knowledge."

Immanuel Kant

ACKNOWLEDGEMENTS

I am so fortunate to have been guided and trained by my wonderful supervisors Ryota Kanai and Anil Seth. Both have shown seemingly endless patience yet have always been encouraging, and have opened great opportunities for me. Perhaps most importantly, they have always been a pleasure and privilege to work with, both intellectually and socially. Rufin VanRullen was a fantastic collaborator on Chapter 5 and I am so grateful for him, and his great team at CERCO, for training me in EEG and having me stay in the lab. Adam Barrett was a great collaborator on the signal detection work, and was especially helpful in the last few weeks of writing this thesis. Andy Mealor deserves special thanks for telling me to add confidence ratings into my paradigm in my first week of the PhD. Who knows what this thesis would have been without that advice.

It has been such a pleasure being at the Sackler Centre, working with great researchers and lovely people. Special thank you to Acer Chang, who has been a source of great emotional and intellectual support and a wonderful PhD-sibling. Both at Sackler and in Psychology, my office mates have always been great people to discuss ideas with, troubleshoot or let just off steam (i.e. share internet tat with): all the people in 1C2, Sam (extra thank you for all the fMRI support), Jeremy, Sarah, Chrissie, Laura, James, Gemma, Michael, Georgina, Acer again, and Marta. To my wonderful friends: Anne, Becky Grist, Bojana, Catriona, Patrick, Gemma, Dan, Josh, Karina, Lucy, Sarah and Suzy.

I am so lucky to have Andy, my incredible partner, best friend (and in the past few weeks especially, thesis mentor), who is amazing in every way. By no means last, to my amazing parents who have been the most supportive and loving parents I could ever ask for. And most of all, to the most wonderful sister imaginable, Alix, for her sweetness, kindness, and generosity, for teaching me so much, our fascinating conversations on the mind, for her unparalleled wit, and more than I could ever put into words. Alixi, your thesis would have been extraordinary (much better than this one). The world has missed out: who else could combine Kant and logical paradoxes to conceive “Liar liar, Kant’s on fire”? I am a much richer person, and so privileged, for having such an extraordinary sister.

CONTENTS

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

CHAPTER 7

LIST OF FIGURES

LIST OF TABLES

THESIS SUMMARY

The notion that unconscious Bayesian inference underlies perception is gaining ground. Predictive coding approaches posit that the state of the world is inferred by integrating, at each level of the perceptual hierarchy, top-down prior beliefs about sensory causes and bottom-up prediction errors. In this framework, percepts correspond to a top-down stream of beliefs that best 'explain away' sensory signals. Although such frameworks are gathering empirical support, subjective facets of perception remain unexplained from these perspectives. This thesis combines behavioural, computational and neuroimaging methods to examine how subjective visual confidence can be accounted for in a predictive coding framework.

Experiment one shows that, behaviourally, perceptual expectations about target presence or absence both liberalise confidence thresholds and increase metacognitive accuracy. These results are modelled in a signal detection-theoretic framework as low-level priors shifting the posterior odds of being correct. Using EEG, experiment two reveals that influence of expectations on decision and confidence oscillates with the phase of pre-stimulus alpha oscillations. This means that prior to target onset, both objective and subjective decisions have been rhythmically biased by the periodic recruitment of expectations to visual areas. Using fMRI, experiment three shows that in the post-stimulus period, expectations and sensory signals are integrated into confidence judgements in right inferior frontal gyrus (rIFG). Furthermore, this process recruits orbitofrontal cortex and bilateral frontal pole, which represent top-down influences, and occipital lobe, which represents bottom-up signals. Together, these results suggest that expectations shape subjective confidence by biasing the posterior probability of the perceptual belief.

1

INTRODUCTION

1.1 Overview

Accompanying our perceptual content is a sense of confidence in what we see. Sometimes our perceptual content is clear, and we feel able to identify the source of our sensory signals. However under sensory uncertainty, for example in the dark or when looking out the corner of our eye, we may become unsure. Subjective perceptual confidence is an important facet of our visual experience, that often reflects our conscious content (Kanai, Walsh, & Tseng, 2010; Sandberg, Timmermans, Overgaard, & Cleeremans, 2010; Seth, Dienes, Cleeremans, Overgaard, & Pessoa, 2008), but that also indicates an ability to evaluate our own judgements.

In many cases we are able to accurately judge the correctness of our perceptual inferences, that is, we demonstrate reasonably high *metacognitive accuracy* (Fleming & Dolan, 2012; Kentridge & Heywood, 2000; Yeung & Summerfield, 2012). We know when we know, and we know when we do not, meaning that our confidence correlates with our perceptual performance. However, confidence is subject to specific biases, such as a systematic underestimation of uncertainty in our environment (Zylberberg, Roelfsema, & Sigman, 2014) and a tendency to avoid evidence for unselected perceptual inferences (Maniscalco, Peters, & Lau, 2016; Zylberberg, Barttfeld, & Sigman, 2012). Despite such biases, little is known about how non-sensory influences shape our sense of confidence.

For objective perception, research increasingly points to a more complicated picture than simple feature extraction, revealing that non-sensory influences such as motivation and beliefs exert powerful, shaping effects. There is now a wealth of evidence showing that perceptual prior expectations about the probable causes of sensation bias perceptual inference, and are associated with a suppression of ERP and BOLD activity (den Ouden, Kok, & de Lange, 2012; Gilbert & Li, 2013; Summerfield & de Lange, 2014). These influences, of “seeing what we believe”, can be formulated in Bayesian terms. Bayesian frameworks propose that perception can be modelled as an integration of prior expectations and sensory data, where the ‘winning’ inference is that with the highest posterior probability (see section 1.2.2). In this way, we will perceive that which best explains our incoming sensory data. ‘Predictive coding’ frameworks extend this notion, proposing that the brain refines and shapes top-down predictions until most of the bottom-up sensory data has been explained away (for a review, see Spratling, 2016).

Predictive frameworks are increasingly being applied to high-level cognition, for example cognitive control (Pezzulo, 2012), theory of mind (Koster-Hale & Saxe, 2013) and sense of agency (Friston, 2014). These frameworks are even being extended to try and explain conditions such as autism (Van de Cruys et al., 2014) and schizophrenia (Horga, Schatz, Abi-Dargham, & Peterson, 2014). In many ways this is not surprising: as will be described in Section 2.2, these frameworks are very rich. Remarkably, subjective facets of perception remain largely unexplored from this perspective, despite a considerable body empirical work on objective perceptual decision-making. There has been some work, both theoretical and empirical, that has examined effects of expectation on the contents of visual consciousness (Hohwy, Roepstorff, & Friston, 2008; Jakob Hohwy, 2012; Melloni, Schwiedrzik, Müller, Rodriguez, & Singer, 2011), on sense of presence (Seth, Suzuki, & Critchley, 2011) and even on synaesthetic experience (Seth, 2014b). However, the extension of predictive processing frameworks to the domain of consciousness remains in its infancy. The feeling of confidence that accompanies perceptual judgements has been particularly neglected.

This neglect persists despite confidence being particularly amenable to Bayesian frameworks. Confidence is often conceived as a subjective probability that a decision was correct (Meyniel, Sigman, & Mainen, 2015; Pouget, Drugowitsch, & Kepecs, 2016), and accordingly, it must involve some inference on internal states or sensory representations. Recent theoretical work has suggested that constructing our sense of confidence involves reading out the posterior probability of the perceptual choice made (Meyniel, Sigman, et al., 2015). Alternatively, confidence could be conceived as a meta-decision, in which the probability of the decision having been correct is inferred in a manner analogous to objective decision-making.

This thesis attempts to bridge research on subjective perceptual confidence and on predictive influences on objective decision-making. It will investigate how top-down influences of prior perceptual expectations, controlling for influences of top-down attention, shape the construction of subjective visual confidence. Using a novel paradigm, work here uses visual psychophysics, EEG and fMRI to show that we are more likely to assign high confidence to predicted percepts This process of integrating expectations into confidence recruits both sensory and frontal brain regions, and begins prior to stimulus onset. Therefore, the work here reveals that predictive processing frameworks can be naturally extended to the domain of subjective decision-making.

This Chapter will briefly introduce perceptual decision-making and metacognition, with the goal of outlining underlying principles of the work in this thesis and prefacing the subsequent Chapters.

1.2 PERCEPTION AS A DECISION

Perception is often regarded as decision-making, in that the processing of sensory signals leads to changes in the evidence for one perceptual inference – or decision – over another. In common to many perceptual decision-making (PDM) models is the notion of a decision threshold, which determines the point at which enough evidence has been accumulated for alternatives to be distinguished. On this account, top-down effects of expectation may push decisions towards one or another alternative by reducing the evidence required for selection. The specifics of these models do not form the focus of this thesis, but they motivate the theoretical foundation of this work. Accordingly, this section will give a broad overview of perceptual decision-making models and decision theory.

1.2.1 Serial sampling

Normative models describe PDM as a process of serial sampling evidence accumulation, whereby noisy evidence accumulates linearly towards a decision bound, and a decision is made when either the relative evidence (diffusion models, for reviews see Forstmann, Ratcliff, & Wagenmakers, 2016; Ratcliff, Smith, Brown, & McKoon, 2016) or absolute evidence (race models, e.g. Brown & Heathcote, 2008; Vickers, 1979) reaches the bound. These are illustrated in figure 1.1.

These serial sampling frameworks model the accumulated evidence over time, stating that at each time point, the change in evidence is equal to some accumulated evidence and some noise. Evidence accumulation begins at the baseline level $x(0) = x_0$, where x_0 is taken to be zero in the absence of prior

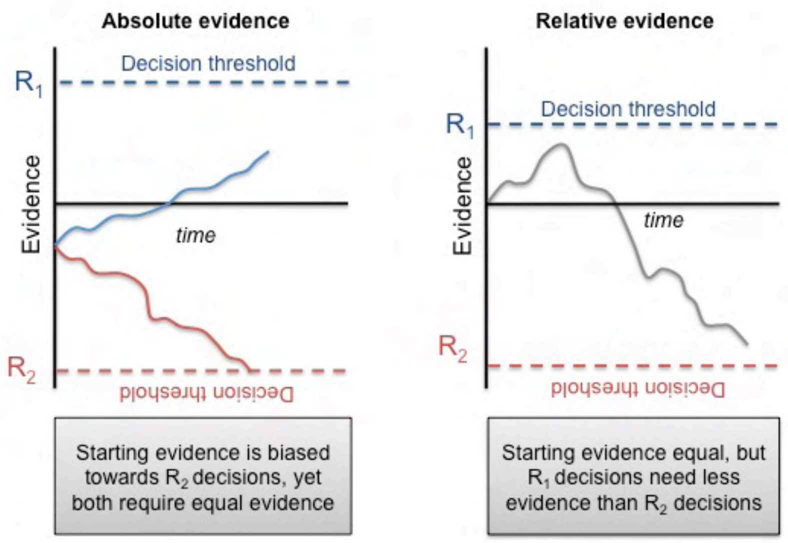
Absolute evidence
Decision threshold
R_1
Evidence
time
R_2
Decision threshold
Starting evidence is biased towards R_2 decisions, yet both require equal evidence
Relative evidence
Decision threshold
R_1
Evidence
time
R_2
Decision threshold
Starting evidence equal, but R_1 decisions need less evidence than R_2 decisions

Figure 1.1 Evidence accumulation

On the left, noisy evidence is accumulated to a decision threshold, or bound, separately for each response type. Whichever accumulator reaches the bound first is selected as the decision. On the right, the relative evidence between the two choices is accumulated. The starting point of evidence accumulation may be biased towards one of the decisions (see left), or the thresholds might be unequal, meaning that one choice needs less evidence than the other for selection (see right).

information. Race models are defined in the same way, except that the evidence for each choice is considered separately, and is accumulated under distinct drift rates with distinct starting points.

In these models, top-down influences could be modelled as a change in x_0, but also as a change in the drift rate, or as a change in the decision bounds, which determine the point at which evidence accumulation should halt (Summerfield & Egner, 2009). Alternatively, priors could simply add noise to the evidence

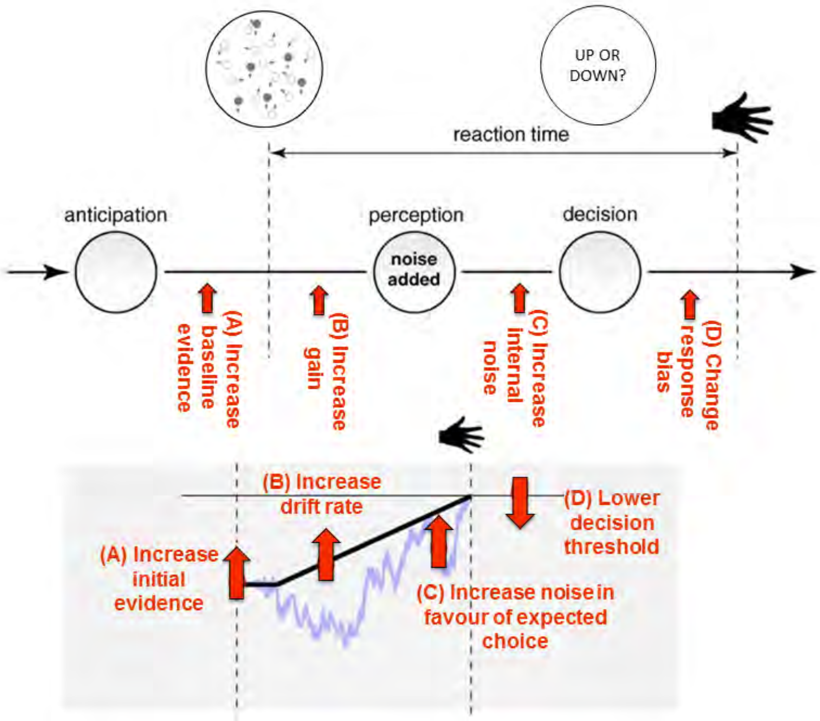
UP OR DOWN?
reaction time
anticipation
perception
noise added
decision
(A) Increase baseline evidence
(B) Increase gain
(C) Increase internal noise
(D) Change response bias
(B) Increase drift rate
(A) Increase initial evidence
(D) Lower decision threshold
(C) Increase noise in favour of expected choice

Figure 1.2 Effect of priors in serial sampling frameworks

Here, the participant has to determine the mean direction of motion (‘up’ or ‘down’) in a random dot kinematogram (RDK). There are four points in the decision-making process at which priors could shape decision-making: (A) The baseline evidence could increase for the choice with higher prior probability. This would occur in the anticipatory stage, prior to stimulus onset and is modelled as a change in initial evidence; (B) Following stimulus onset, expectations could change the gain, or signal to noise ratio, which would be modelled as an increase in drift rate for the expected choice; (C) Internal noise could be added, artificially pushing evidence accumulation in favour of the expected choice; (D) Expectations could alter response biases, which would be modelled as a lower decision threshold for expected choices. Adapted from Summerfield & Egner (2010).

accumulation process so that ‘false’ evidence in favour of the preferred option is collected. Thus, these models offer four key targets at which top-down influences could act, and this is illustrated in figure 1.2.

1.2.2 STATISTICAL DECISION MAKING

An alternative formulation for PDM models describes perception in Bayesian terms. Here, for sensory evidence *x*, evidence for one choice, P(C_1|*x*), is compared against that for the alternative, P(C_2|*x*), in the form of the log-likelihood ratio, LLR. This is defined as:

$$LLR = \log\left(\frac{P(C_1|x)}{P(C_2|x)}\right).$$

When sensory evidence for C_1 exceeds that for C_2, R will take positive values whereas in the opposite case, R will be negative. Accordingly, we can determine a decision rule that determines which choice *C* should be made given the evidence:

$$C = \begin{cases} C_1 & if\ R > 0 \\ C_2 & if\ R < 0 \end{cases}$$

In this framework the decision rule does not usually change in the presence of expectations. Rather, to model expectations we simply consider prior odds of C_1 relative to C_2 and add this to the LLR. Now, LLR becomes

$$LLR = \log\left(\frac{P(C_1|x)}{P(C_2|x)}\right) + \log\left(\frac{P(C_1)}{P(C_2)}\right)$$

The addition of these prior odds pushes *R* towards positive or negative values and thus towards the response that the prior favours. By Bayes' rule, LLR is now equivalent to the *posterior odds ratio*: the relative probability of the two decisions, conditioned upon the evidence and prior odds.

We can also incorporate uncertainty into Bayesian decision theory. Under uncertainty, both the sensory data (modelled as the likelihood) and the prior are represented as Gaussians. The degree of uncertainty about each of these variables can be represented as the respective variances of these distributions. When combined, the posterior will also be Gaussian, with mean

$$\mu_{Posterior} = \frac{\mu_{Likelihood}\ \sigma^2_{Prior} + \ \mu_{Prior}\ \sigma^2_{Likelihood}}{\sigma^2_{Likelihood} + \ \sigma^2_{Prior}}$$

Here, the mean of the likelihood and the mean of the prior are weighted by the variance of the other variable. This is important, because the posterior odds (and therefore the decision) will be weighted according to relative sensory uncertainty. High sensory uncertainty will push the posterior mean towards the prior mean, whereas low sensory uncertainty will push it towards the mean of the likelihood. Accordingly, Bayesian decision theory predicts that decisions will be based upon expectations more when sensory uncertainty is high, and less so when sensory uncertainty is low. Similarly, if there is high uncertainty about the prior, for example if the environment is volatile and rules frequently change, expectations will carry greater weight upon the decision.

1.3 DECISION CONFIDENCE

In the decision-making literature – in perception, or learning and memory - participants are generally asked to make a forced choice about stimuli presented to them: ‘is this word old or new?’ (a yes/no task), or ‘was the target presented on the left or on the right?’ (2AFC). Answers to these decisions tell us about how stimuli are processed and reported, but these decisions are also accompanied by a sense of subjective confidence in our choice. These confidence judgements represent the subject’s probabilistic belief that they have made the correct decision, and usually correlate with decision accuracy (Grimaldi, Lau, & Basso, 2015). This tells us that we are not just good perceptual decision-makers, but are able to evaluate those decisions appropriately: we ‘know when we know’.

Understanding how these confidence judgements are constructed informs us about how we make and evaluate our decisions, for example about how we are able to learn from our mistakes (Yeung & Summerfield, 2012). Determinants of high confidence may also help us understand the way in which knowledge or perceptual content becomes conscious (Seth et al., 2008; Wierzchoń, Paulewicz, Asanowicz, Timmermans, & Cleeremans, 2014). Confidence has post-decisional benefits as well. For example, confidence can guide perceptual learning (Guggenmos, Wilbertz, Hebart, & Sterzer, 2016), it may act as a 'common currency' between perceptual inferences, facilitating the integration of several information sources (de Gardelle, Le Corre, & Mamassian, 2016; de Gardelle & Mamassian, 2014), and communicating decision confidence with peers improves group decision-making (Bahrami et al., 2010; Zarnoth & Sniezek, 1997).

The notion that we can 'know that we know' – that metacognition is intimately tied to knowledge – has a long tradition in philosophy. It can be traced back to Aristotle, who posited that memory requires reflection or conception (Sorabji & Aristote, 1972), and Augustine, who proposed that the mind continually reflects upon itself to understand and develop (Perricone, 2011). In psychology, introspection – the process of evaluating or reflecting upon one's internal states – was revisited by Peirce and Jastrow in 1884, who revealed that decisions on stimuli associated with very low signal strength (small changes in weight) can be above chance, yet be reported with no confidence (Peirce & Jastrow, 1884). They claimed, as we still do, that decision confidence reveals current states of knowledge, concluding with the observation that sensation can be unconscious:

> "[This set of findings] *gives new reason for believing that we gather what is passing in one another's minds in large measure from sensations so faint that we are not fairly aware of having them, and can give no account of how we reach our conclusions about such matters.*"

Though now we know that confidence is not purely a function of signal intensity (see Chapter 2), Peirce and Jastrow crucially showed that accuracy and confidence can dissociate, revealing that conscious knowledge or awareness seems to be a 'privileged' state that not all sensations are granted.

1.5 THESIS OVERVIEW

This thesis will empirically address the question of whether and how perceptual prior expectations shape subjective judgements of perceptual confidence using visual psychophysics, signal detection theory, scalp electroencephalography (EEG) and functional magnetic resonance imaging (fMRI).

Chapter 2 presents an overview of the theoretical and empirical literature on Bayesian brain frameworks and subjective confidence judgements that motivates this thesis. The ways in which confidence and metacognition can be studied empirically are detailed in Chapter 3. Chapter 4 presents a comprehensive analysis of behavioural data, showing that prior expectations increase subjective confidence and improve metacognitive sensitivity. Furthermore, it presents a Bayesian signal detection theoretic framework that accounts for these results. Using EEG, Chapter 5 shows that perceptual priors begin to bias objective and subjective judgements prior to stimulus onset. Using fMRI, Chapter 6 combines general linear modelling (GLM), psychophysiological interaction analysis (PPI) and voxel-based morphometry (VBM) to reveal a functional network in which confidence is shaped by perceptual priors. Chapter 7 consolidates these findings with respect to the literature and presents a simple hierarchical Bayesian scheme that offers one plausible solution for how confidence could be modelled in a hierarchical predictive coding framework.

2

LITERATURE REVIEW

2.1 Overview

What we perceive is not only a function of the external world, but also of what is expected by the observer (Gilbert & Li, 2013). Such expectations may be in the form of contextual information, or on previous experience in similar environments. One might imagine that these top-down influences act at late response stages of perceptual processing, however evidence increasingly points to expectations exerting effects at very early stages (Rauss, Schwartz, & Pourtois, 2011). This means that these top-down influences cannot simply arise from post-hoc reasoning, but rather, that perceptual inference is constrained by prior beliefs. These prior beliefs can pertain to knowledge of natural scene statistics, for example, that luminance changes at the edges of an object. Alternatively, they can reflect the probability of a given object in the current environment, or even beliefs passed down by evolution. This influence of prior beliefs on perception has led to growing interest in formulating perception in Bayesian terms (see Friston, 2012b; Knill & Pouget, 2004). Here, perceptual content corresponds to the hypothesised cause of sensation that is most probable, given the sensory signals and priors. Under these frameworks, probable causes are more likely to be selected, and unlikely causes are more likely to be suppressed. These ‘Bayesian Brain’ frameworks motivate the work presented in this thesis, though the work here neither *directly* tests them, nor depends upon them.

This overview chapter will describe process models that attempt to explain how perception is shaped by top-down influences. Next, it will outline the current state of knowledge on how top-down influences shape perceptual decisions, including the neural substrates underlying subjective decisions. Finally, this section will bring these findings together and give a brief overview of how top-down expectations might shape confidence judgements. This final question forms the topic of the present thesis.

2.2 The Bayesian brain hypothesis

How can the brain infer the state of the world from only indirect, ambiguous sensory signals for which there is no one-to-one mapping between sensation and cause? A shadow, as depicted in figure 2.1A, may appear to have only one plausible sensory cause – a guitar – however there are many alternative causes, such as that depicted in figure 2.1B. We can also perceive, at distinct points in time, multiple perceptual interpretations of the same stimulus. This is made apparent in bistable phenomena such as Ruben's face/vase illusion (fig. 2.1C), which can be perceived either as two faces or as a vase. Yet despite these conflicts, our survival indicates that we are able to infer the causes of our sensations with high degree of accuracy.

Bayesian brain frameworks propose that perception is achieved via Bayesian inference, that is, in a manner consistent with Bayesian statistical decision making. Here, perceptual content is said to be the hypothesised cause of sensation that has maximum posterior probability, given the data and prior beliefs. Sensory signals and prior beliefs are combined into the posterior belief using Bayes’ rule, which states:

$$P(Hypothesis \mid Data) = \frac{P(Hypothesis) \times P(Data \mid Hypothesis)}{P(Data)}$$

However, the probability of the data is just a constant as it is invariant to the hypothesis. Therefore, we get:

$$Posterior \ \propto Prior \times Likelihood.$$

This arises via a process of hypothesis testing, in which decisions are unconsciously guided by learned regularities in the world (Helmholtz, 1860). By simulating the expected probabilistic outcomes (posterior) under each perceptual hypothesis, the most explanatory hypothesis can be determined (Gregory, 1980). Thus, perceptual content arises from this process of iteratively testing sensory signals against hypotheses until as much of the sensory data as possible are ‘explained away’. The most explanatory perceptual hypothesis is that which maximally explains the sensory signals, given context-dependent prior beliefs.

This account assumes that the brain represents internal models that represent both these learned regularities (prior beliefs on sensory causes), and mappings between sensory causes and sensory data (likelihoods). That is, it must represent rules that govern sensory effects. However, in order to calculate the posterior the brain must infer unknown (hidden) causes from known, internally modelled effects, that is, perform ‘backwards’ inference. More formally, this is referred to as inverting the generative model.

If perception is achieved by hypothesis testing, where do the priors come from? How should, and why can, perception arise in this way? These questions are

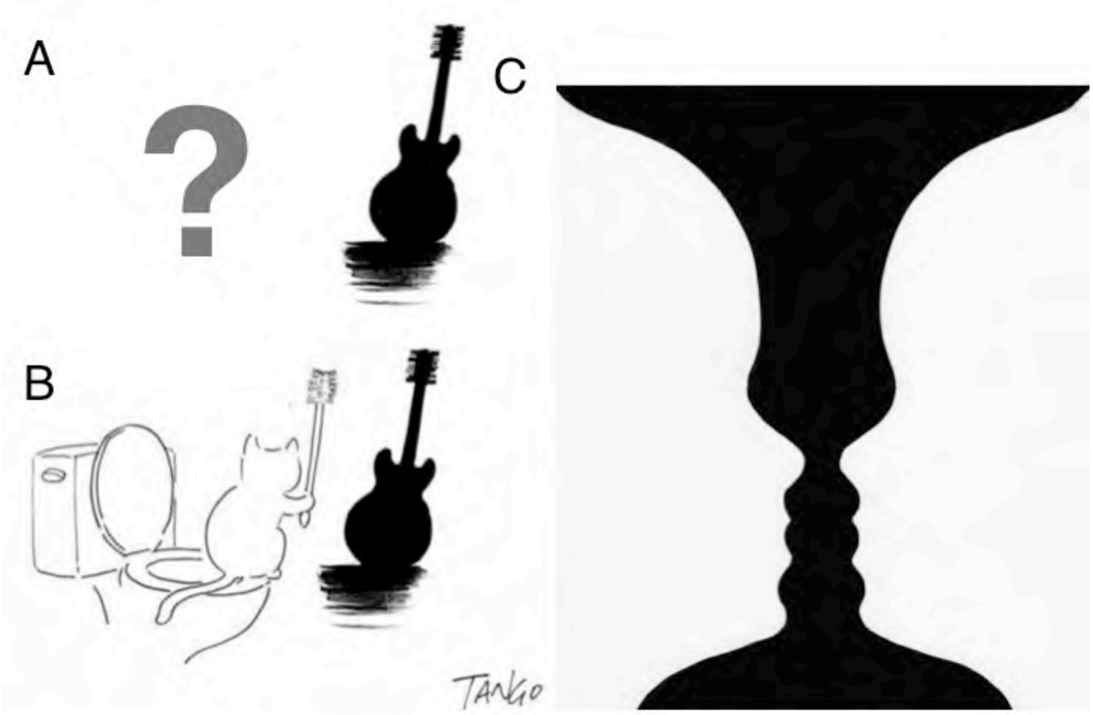

Figure 2.1. Multiple perceptual interpretations

A. The most probable sensory cause of this shadow is a guitar, however as shown in **B.** there are also unlikely sensory causes (a cat holding a toilet brush). This image illustrates how we often constrain our inferences according to prior probabilities.
C. Ruben’s face/vase illusion. This image is considered bi-stable because it has two possible interpretations: the black shape can be perceived as a vase, or the negative space can be perceived as two people facing each other.

arguably best addressed in a biologically plausible fashion in the free-energy principle (FEP; Friston, Adams, Perrinet, & Breakspear, 2012; Friston, 2009, 2010). In FEP, the Bayesian brain hypothesis is a corollary that follows from a unifying explanation of why living organisms operate or behave as they do.

This principle starts with the premise that the goal of any living organism is to maintain its homeostasis, and that this is achieved by maximising the probability that the organism will remain in a small set of possible states (minimising entropy). In order to remain in a small set of possible states, the agent must minimise its long-term average *surprise*. What is surprising depends on the organism. For example, a fish out of water would be in a surprising state for a fish (Friston, 2010), but this would not be surprising for a human.

How can surprise be minimised? FEP recruits the fact that free-energy is an upper-bound on surprise to propose that organisms maintain their homeostasis by minimising free-energy, that is, maximising their internal model evidence. With some assumptions (see 2.2.1), all the organism has to do is minimise the discrepancy between its sensory states and expected sensory causes (beliefs, or expectations). Minimising this discrepancy will minimise free energy, thereby minimising surprise and supporting homeostasis. The discrepancy between sensory states and expectations can be minimised in two ways. One possibility is to act upon the world to change sensation; another is to change internal states by altering beliefs (figure 2.2). More informally, one can think of the brain as a sophisticated but corrupt scientist, who over time tweaks his predictions and manipulates his data in order to consistently support his own larger theory.

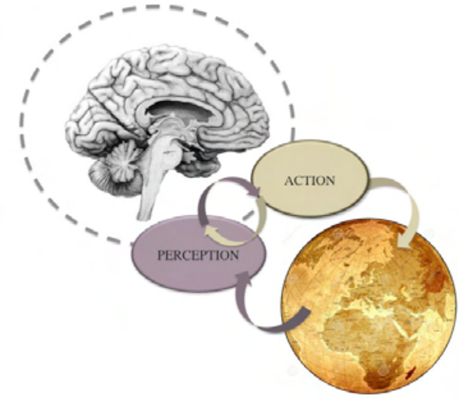

Figure 2.2 The perception-action loop under the free-energy framework
The brain has only indirect access to the world, achieved via action and sensation. The agent can change sensation via action, or change perception by altering its prior beliefs.

2.2.1 Variational Bayes and hierarchical predictive coding

Under this theoretical foundation for *why* the brain is a Bayesian hypothesis tester, we can ask how this might be achieved. How is the posterior belief inferred? The free-energy principle (FEP) relies upon a form of Variational Bayesian inference, implemented by the brain in the form of hierarchical predictive coding (for a technical formulation, see Friston & Kiebel, 2009). Note that there those who are proponents of predictive coding as a functional architecture who do not necessarily subscribe to FEP (e.g. Spratling, 2008).

FEP takes a Variational Bayesian approach to approximating the posterior density *P(cause* | *data).* Here, the posterior is approximated by a recognition density *Q(cause).* The goal is to minimise the discrepancy between P and Q. In perception the data do not change (data change via action); the priors (i.e. the recognition density) must be optimised to minimise the distance between these distributions. FEP assumes that the brain uses the Laplace approximation, meaning that (i) these densities are Gaussian and (ii) the variance of these Gaussians is a function of their mean.

Recall that the brain is thought to use hierarchical generative models. These furnish a stream of empirical priors, that is, conditionally dependent beliefs about sensory causes. In the brain, this maps well onto modular or hierarchical processing. We can imagine priors for edges in V1, for object shape in lateral occipital complex (which are dependent upon edges), and for semantic recognition of the object in fusiform gyrus (which are dependent upon edges and shape). Within this hierarchical scheme, optimising the recognition density entails optimising interdependent priors across the cortical hierarchy. Together, this process gives an approximation of the true posterior. FEP proposes that the brain performs Variational Bayesian approximation by implementing a hierarchical predictive coding scheme with the Laplace approximation (for a review, see Spratling, 2016).

Hierarchical predictive coding is a message-passing algorithm that optimises sufficient statistics in each level of the hierarchy by minimising the discrepancy between top-down priors and bottom-up prediction errors, until all prediction errors converge within some small margin of error (figure 2.3). Prediction error corresponds to the discrepancy between the prior and the data. The idea here is that separate units (in the brain, neurons) will represent priors and prediction errors. Each task-relevant region of the perceptual hierarchy receives top-down priors and bottom-up prediction errors, and from these, estimates the posterior. This posterior forms the prior for the level below, and any unexplained data that remains is communicated to the hierarchical level above as prediction error, to be 'explained-away' at increasing levels of abstraction. In parallel, prediction error will be used to update the generative model such that predictions are optimised in the longer-term as well.

Thus, predictive coding assumes that perceptual inference is achieved via a cascade of reciprocal exchanges of predictions and prediction errors, which minimises prediction error across the cortical hierarchy. In this way, (some) priors are *empirical*, meaning that they do not have an origin, but rather are constrained by inferences across the brain, as well as across timescales.

2.2.2 Attention as precision weighting

Crucial to Bayesian statistical decision theory, independently of neural implementation, is the notion of precision. Here, priors and data can be represented as Gaussian probability distributions, which means that they will be associated with a precision (variance) representing their reliability. In volatile environments, where governing rules frequently change, or if these rules are unclear (you are in a novel environment, say) it would be unwise for inference to place too much weight on the prior belief. Its precision will be low. Similarly, if sensory uncertainty is expected to be high – say it is a dark night and you are outside – it would be unwise for inference to place too much weight on these sensory signals. Bayes rule ensures that priors and likelihoods (sensory signals) will be combined with respect to their precisions (see section 1.2.2 and figure 2.4).

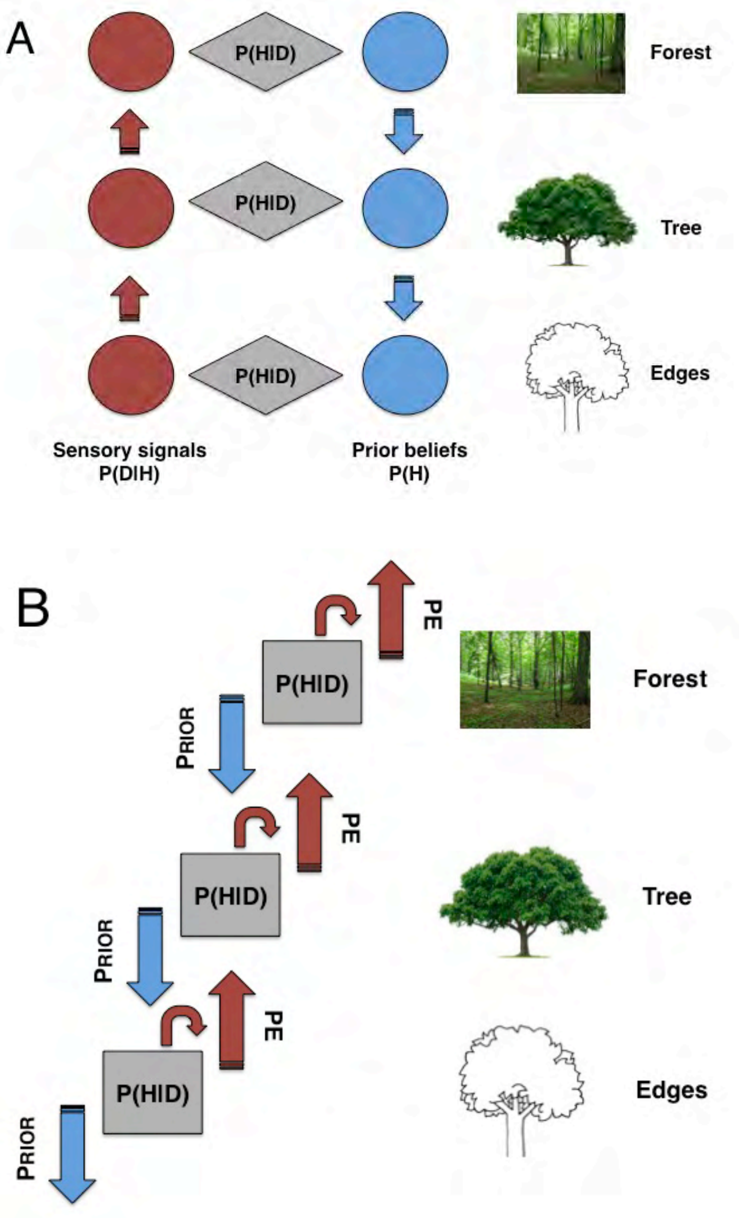

Figure 2.3 Hierarchical inference.

(A) Hierarchical Bayesian inference of a forest scene. Top-down prior beliefs (blue) and bottom-up sensory signals (red) are combined at each hierarchical level into a posterior belief using Bayes rule. At higher levels of the hierarchy, inference corresponds to more abstract or global representations. Here, inference moves from perception of edges, to perception of objects, to perception of the wider environment.
(B) Hierarchical predictive coding. This is similar to hierarchical Bayesian inference. The crucial difference is that the bottom-up information is prediction error (PE), not sensory signals. At each level of the hierarchy, top-down priors are received from the level above and bottom-up PEs are received from the level below (and in lateral connections).The inference at each stage constrains that at the level below by becoming the prior. It also constrains the level above, by feeding-forward remaining PE.

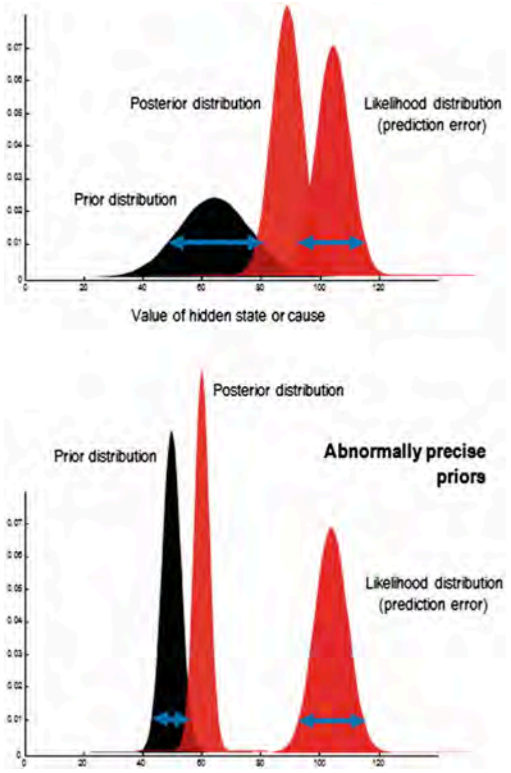

Figure 2.4 Combining priors and likelihoods

The top panel shows that when the prior distribution is wide relative to the likelihood function, i.e. there is *high prior uncertainty,* the posterior will be weighted more by (i.e. look more similar to) the likelihood. In the bottom panel the prior is very precise. Accordingly, the posterior is weighted more by the prior than the likelihood. Picture taken from (Edwards, Adams, Brown, Pareés, & Friston, 2012).

So, within the free-energy principle what is the role of top-down attention, for example selective attention? One intriguing proposal is that attention is the process of adjusting precision weightings according to top-down goals and motivations, like tuning a radio or tweaking an amplifier (Clark, 2015; Kanai, Komura, Shipp, & Friston, 2015). In this way, the relative contributions of top-down priors and bottom-up prediction errors can be adjusted by top-down beliefs on precisions, accentuating the signals that are most relevant to the agent. To illustrate, imagine you are on an Easter egg hunt, and desperate to win. You will be looking hard for anything remotely egg-shaped, but anything irrelevant to the task (even if it is usually motivationally salient), you will try to ignore. Within the free-energy framework, this would be modelled as an expectation that you will win the hunt – that prior represents the motivation. This belief (motivation) shapes second-order prior beliefs on the precisions in stimulus-relevant sensory and association cortices. These beliefs are that egg-coloured or egg-shaped objects have high precision, and anything else has low precision. The result of these beliefs is that information that is consistent with the sensory causes that are relevant to your goals – Easter eggs – will be prioritised, and all other information will be suppressed. In figure 2.4, this would mean changing the variance of the prior and likelihood functions, so that the posterior distribution is biased towards the more reliable evidence.

This concept has received only preliminary and circumstantial support, primarily from modelling work (Brown & Friston, 2012; Feldman & Friston, 2010; Kanai et al., 2015; Vossel et al., 2014), however its utility is primarily in the explanatory power it affords. Under this framework, predictive processing governs not only perception, but a host of high-level cognitive processes (Clark, 2015) such as motivation and value (Friston & Ao, 2012; Friston et al., 2015).

2.2.3 Neural architecture underlying a Bayesian brain?

So far, it has been proposed that interactions between the brain and its environment are mediated by perception and action, such that perception optimises predictions and action changes the sensory inputs. Together these processes maximise an agent's internal model evidence. As yet, much of the neuroanatomical and neurobiological evidence for such a framework is circumstantial, though plausible neural architectures have been proposed in considerable detail (Bastos et al., 2012; Friston & Kiebel, 2009a; Kanai et al., 2015). One prediction of these architectures is that priors are communicated in beta band oscillations, precision-weighted prediction errors are communicated in the gamma band, and precision-weighted priors are communicated in the alpha band. This hypothesis has received recent support from model-based analyses on human electrocorticography (eCoG) data (Sedley et al., 2016).

However, evidence for the existence of neurons that signal perceptual prediction error remains elusive. Their existence is not implausible, given evidence for the existence of reward prediction error neurons in midbrain (Bayer & Glimcher, 2005; Nakahara, Itoh, Kawagoe, Takikawa, & Hikosaka, 2004; Schultz & Dickinson, 2000), but this assumption – of prediction error neurons – renders FEP particularly controversial. So, what are alternative neural codes for Bayesian inference, if not priors and prediction errors? One possibility is that neurons, at least in visual cortex, represent the entire posterior probability distributions (Fiser, Berkes, Orbán, & Lengyel, 2010; Hoyer & Hyvarinen, 2003; Paulin, 2005). This idea is lent circumstantial support from the fascinating finding that visual cortex neurons vary in the degree to which their activity is correlated with their neighbours' (Okun et al., 2015). This is what we would expect if probability distributions were represented at the group level: most neurons signal similar information (are close to the peak), while fewer neurons signal deviant information (are close to the tails). Alternatively, some frameworks propose that neurons encode the sufficient statistics, that is, the mean and standard deviation of the posterior density functions (Knill & Pouget, 2004), or that their activity reflects the inference itself (Lee & Mumford, 2003).

The ideas summarised in this section – of the brain as a Bayesian inference machine, and particularly the free-energy principle – are not argued to be 'true', and are not presented in order to introduce work aiming to lend support to this framework. Rather, this framework offers us a set of unifying principles under which to consider cognition, and allows us to conceptualise processes from perception to attention, expectation, value, motivation and more, in terms of just a small set of concepts common to all function: hierarchy, priors, prediction errors and precision. This set of unifying principles, which, in its free-energy formulation, has been proposed as a unifying theory of the brain (Friston, 2010), motivates a key question which forms the focus of this thesis: if all brain function can be described in terms of priors and precision, consciousness should be explicable in these terms too. Thus, this thesis focuses on the role of priors in one particular facet of consciousness that is relatively amenable to quantification: subjective perceptual confidence.

2.3 ATTENTION AND EXPECTATION SHAPE PERCEPTION

The way in which priors and precisions should shape perceptual content is largely consistent with empirical work on expectations and top-down attention. These top-down influences are usually studied separately, yet many paradigms thought to manipulate one may additionally manipulate the other. For example, Posner cueing (see figure 2.5) should divert attention towards the location at which the stimulus should appear. However, while this paradigm is used to look at *attentional* effects on reaction time, the cue concurrently induces a probabilistic *expectation* that the probe will appear in a particular location (Summerfield & Egner, 2009). Therefore, before considering how attention and expectation might shape confidence, it is important to determine how they shape objective perceptual judgements.

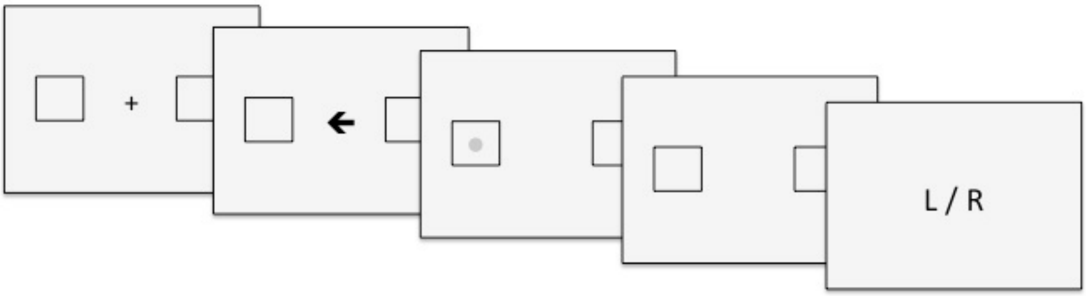

Figure 2.5 Posner cueing paradigm

Here, a spatial cue (usually with 75% validity) is presented before stimulus onset so that covert attention can be allocated to the task-relevant spatial location. A probe is then presented, either in the cued or uncued location. Here, the probe appears in the cued location. After the probe has been removed from the screen participants are asked to report the side on which the target appeared as fast as possible. The critical comparison in this task is reaction time differences between cued (attended) and uncued (unattended) probes.

2.3.1 TOP-DOWN ATTENTION

The multiple forms of attention are generally divided into two categories: one is stimulus-driven attention, often termed *attentional capture*; the other is top-down attention, which is driven by the agent’s goals or desires, and is under volitional control (Theeuwes, 2010).

Behaviourally, top-down attention improves the quality of information, adjusting the signal to noise ratio for relevant targets by determining features that should receive priority versus those that should be suppressed (Knudsen, 2007). Accordingly, attention increases the responsiveness of early visual cortex neurons to task-relevant signals (Martinez-Trujillo & Treue, 2004; McAdams & Maunsell, 2000; Saenz, Buracas, & Boynton, 2002). One popular idea is that top-down attention acts after a feedforward ‘sweep’ so that goals and motivation can target appropriate regions of the visual field (Bar et al., 2006; van Gaal & Lamme, 2012) and bias competition in favour of more important targets (Beck & Kastner, 2009; Desimone & Duncan, 1995; Hickey & Theeuwes, 2011).

While attention tends to target signals in sensory and parietal cortices, its deployment seems to originate in the dorsal attention network: the frontal areas frontal eye fields (FEF) and intraparietal sulcus (IPS, see figure 2.6). Both FEF and IPS are candidate regions for representing salience or priority ‘maps’, representing regions of space according to their behavioural relevance, that can be read by perceptual or oculormotor regions (Serences & Boynton, 2007).

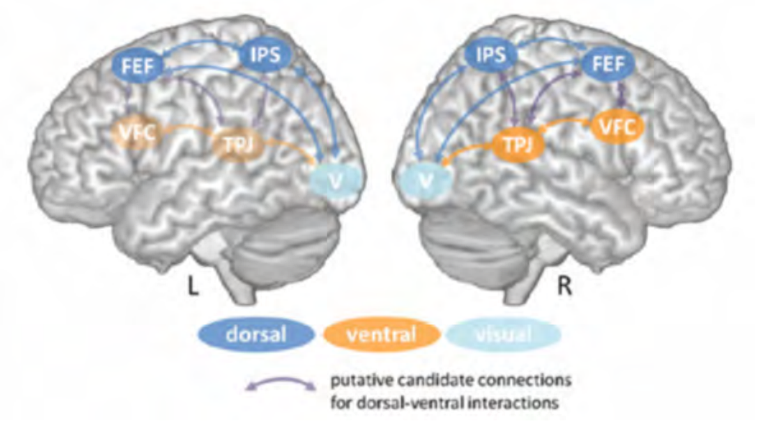
FEF
IPS
VFC
TPJ
V
L
IPS
FEF
VFC
TPJ
V
R
dorsal
ventral
visual
putative candidate connections
for dorsal-ventral interactions

Figure 2.6. Dorsal and ventral attention networks.

Both networks are connected with visual cortex (V). The dorsal attention network consists of frontal eye fields (FEF) and intraparietal sulcus, (IPS), whereas the ventral network consists of ventral frontal cortex (VFC) and temporoparietal junction (TPJ). Figure from Vossel, Geng & Fink (2014).

Stimulus-driven attentional processes, particularly the detection of behaviourally relevant but *unexpected* targets, have been associated with a second network - the ventral attention network. This consists of temporoparietal junction (TPJ) and ventral prefrontal cortex (VPF; Corbetta & Shulman, 2002). These two networks most likely interact, deploying attention according to both top-down and bottom-up influences (Vossel, Geng, et al., 2014).

In summary, top-down attention is conceived as a prioritisation process, which improves the quality of goal-relevant signals. Predictive processing frameworks model these effects as top-down modulation of precision-weighting, adjusting neural gain according to those signals that are of most importance.

2.3.2 Prior expectation

Top-down attention and prior expectations are conceptually distinct: while attention is allocated to signals or processes on the basis of priority or relevance, expectations tell us about what those signals are most likely to be (Summerfield & Egner, 2016). Their effect on perception is undoubtedly powerful (Gilbert & Li, 2013; Rauss et al., 2011; Sarter, Givens, & Bruno, 2001; Theeuwes, 2010). When a target is expected, for example it is predicted by a cue, because it is associated with the context, or because it occurs frequently, reaction times substantially decrease (Coste, Sadaghiani, Friston, & Kleinschmidt, 2011; Eickhoff, Pomjanski, Jakobs, Zilles, & Langner, 2011; Jaramillo & Zador, 2011), even after controlling for response preparation or anticipation (Umbach, Schwager, Frensch, & Gaschler, 2012). These findings are often interpreted as evidence that priors shape evidence accumulation, either by changing the initial evidence for more probable inferences, or by lowering decision thresholds for expectation-congruent choices (see section 1.2.1).

Whether expectations about perceptual content also increase perceptual sensitivity is a matter of debate. Accuracy and sensitivity measures tend to remain unaffected by expectation manipulations (e.g. Kok, Brouwer, van Gerven, & de Lange, 2013; Kok, Rahnev, Jehee, Lau, & de Lange, 2011;

Morales et al., 2015), but reverse correlation analyses suggest that expectations may increase signal to noise ratio (Cheadle, Egner, Wyart, Wu, & Summerfield, 2015; Wyart, Nobre, & Summerfield, 2012). These positive results are consistent with a role of expectations in facilitating the evidence accumulation process, for example, by adjusting neural gain, or precision, such that the neural response to probable features is amplified (Summerfield & Egner, 2016). While the role of expectations in perceptual sensitivity remains unclear, we know that expectations can bias decisions in favour of the more probable sensory cause, consistent with Bayesian principles (den Ouden, Kok, & de Lange, 2012; Summerfield & de Lange, 2014; Summerfield & Egner, 2009).

Indeed, there is now a wealth of evidence showing that a lot of perceptual decision-making is Bayes-optimal, in the sense that sources of information are combined in such a way that the influence of the most reliable information is maximised (see Chapter 1). For example, de Gardelle and Summerfield (2011) presented participants with a circular array of hues and asked participants to report the average hue. They factorially manipulated the mean (i.e. expected) hue and the variability of hues, and found that high means and low variances independently improved accuracy and sped decisions on task-relevant feature dimensions. Crucially, hues that were closer to the mean hue carried greater weight in participants' averaging than outliers. This appropriate use of summary statistics when down-weighting outliers can be interpreted as optimal Bayesian inference (Feldman, Griffiths, & Morgan, 2009). Bayes-optimal incorporation of priors has been evidenced in a wide range of domains, including the estimation of motion trajectories (Körding & Wolpert, 2004), the application of force (Koerding, Ku, & Wolpert, 2004; Singh & Scott, 2003), object perception (Kersten, Mamassian, & Yuille, 2004), and cross-modal cue integration, such as across the visual and haptic domains (Ernst & Banks, 2002).

The Bayes optimal observer provides a benchmark against which human performance can be compared. When behaviour is fit well by Bayesian decision theory, minimal conditions for the brain being Bayesian, or engaging in predictive processing, have been met. However, non-Bayesian or suboptimal inference does not necessarily refute the Bayesian brain hypothesis. Rather, suboptimal inference may reveal constrains on the cognitive system, or reflect non-trivial priors and/or utility functions that we hold but that are not modelled, such as learned preferences and goals. For example, the size-weight illusion, where larger items are perceived as lighter than smaller items, is considered ‘anti-Bayesian’, because inference goes in the opposite direction from the prior expectation that larger objects are heavier (Brayanov & Smith, 2010). The potential to accommodate anti-Bayesian phenomena in Bayesian schemes has led to criticisms that the framework is unfalsifiable (e.g. Bowers & Davis, 2012). One recent proposal has put forward the idea that the perceptual system does behave according to Bayesian principles, but uses an efficient coding scheme that maximises mutual information between the stimulus and the sensory representation. The authors show that under certain loss functions and when there is internal noise, this efficient coding scheme can lead to skewed likelihood functions, resulting in posteriors that are pushed away from the prior: ‘anti-Bayesian’ effects (Wei & Stocker, 2015). This model illustrates how apparently suboptimal behaviour can arise from constraints on the cognitive system – here, the requirement of data compression.

The notion that probabilistic information shapes perception is not in doubt (Firestone & Scholl, 2015), however it remains unclear how this occurs. For example, how are expectations encoded? How are expectations compared to sensory signals, and at what stage in the decision-making stream do they act? A key question addressed in this thesis that also remains unstudied is: to what extent do these influences carry over into subjective facets of perception? If they do carry over, how does this occur?

2.3.2.1 Neural correlates of prediction

A key region implicated in predicting the sensory consequences of an action is cerebellum (Blackwood et al., 2004; Paulin, 2005). Cerebellar activity correlates with the discrepancy between predicted and actual sensory consequences of action (Blakemore, Frith, & Wolpert, 2001). Accordingly, monkeys with cerebellar legions show impairments in predicting future states on the basis of current motor states (Ebner & Pasalar, 2008), and human patients with cerebellar legions are impaired in their ability to update predictive models about sensory consequences of action (Roth, Synofzik, & Lindner, 2013; Synofzik, Lindner, & Thier, 2008).

With regard to how these priors shape sensory processing, it has been hypothesised that priors are communicated across cortical areas via oscillatory activity in either the alpha or beta bands (Arnal & Giraud, 2012; Andre M Bastos et al., 2012; Engel, Fries, & Singer, 2001; Engel & Fries, 2010). Alpha-band oscillations tend to be associated with top-down influences in the perceptual domain (Klimesch, Sauseng, & Hanslmayr, 2007; Mayer, Schwiedrzik, Wibral, Singer, & Melloni, 2015; Zumer, Scheeringa, Schoffelen, Norris, & Jensen, 2014), whereas these effects move to the beta-band in motor tasks (de Lange, Rahnev, Donner, & Lau, 2013; Engel & Fries, 2010; van Ede, Jensen, & Maris, 2010). In accordance with predictive frameworks, there is accumulating evidence for a crucial role of anticipatory (i.e. prestimulus) activity in these frequency bands associated with feedback signalling (van Kerkoerle et al., 2014), in representing expectations of both ‘what’ (Mayer et al., 2015) and ‘when’ (Samaha, Bauer, Cimaroli, & Postle, 2015), as well as in updating rules in accordance with cues (Cooper, Darriba, Karayanidis, & Barceló, 2016). Thus, 8-20Hz neural oscillations are a strong candidate mechanism for the communication of priors.

It also seems apparent that these priors are communicated to, if not represented in, sensory areas. When a portion of the visual field is occluded, the content of that occluded portion can be predicted from its surrounding context, such that in V1 non-stimulated regions are ‘filled-in’ (Petro, Vizioli, & Muckli, 2014; Smith & Muckli, 2010) by feedback signals (Morgan, Petro, & Muckli, 2016; Muckli et al., 2015). In V1, predicted stimuli are also more easily decoded, suggesting that their representation might be more precise (Kok, Jehee, & de Lange, 2012). However, there is relatively little convergent evidence for how and where (and whether) priors are encoded, which is unsurprising given that priors will be specific to the task and stimuli at hand. It is therefore unclear whether findings from under one design would be generalisable to another. To illustrate, in simple low-level perceptual tasks, middle occipital gyri and fusiform gyri may represent relevant prior information (Summerfield & Koechlin, 2008), whereas priors for visual categorisation might be represented in medial frontal gyrus activity (Hansen, Hillenbrand, & Ungerleider, 2012). Orbitofrontal cortex may represent delusional perceptual beliefs (Schmack et al., 2013) and expected reward, while prior beliefs about expected reward may be represented in striatum (d’Acremont, Schultz, & Bossaerts, 2013).

Brain regions that are thought to represent the prior may instead represent the posterior, as these two quantities are tightly associated. To address this concern, Ting and colleagues (2015) used a value-based learning paradigm, in which participants had to estimate the signalled value on each trial by integrating information from two cues: one which reflected the prior probability of reward and the other, the likelihood. Using model-based fMRI, they found posterior probabilities represented in medial prefrontal cortex (mPFC). Furthermore, mPFC also represented prior and likelihood. These results render mPFC a plausible candidate for computing the posterior belief in this task. On a similar vein, Vilares and colleagues (2012) orthogonally manipulated the prior and the likelihood, and found representations of prior information in putamen, insula and amygdala, of the likelihood in occipital cortex, and of the weighting function of prior and likelihood (precision weighting) in superior mPFC. Shifts in baseline evidence induced by priors were reflected in a fronto-parietal network, including medial frontal gyrus, but also superior and inferior frontal gyri and anterior cingulate. Though these findings exhibit variability in where Bayesian quantities are represented, they suggest that priors are generally represented in frontal regions, particularly mPFC, whereas the likelihood is represented in stimulus-specific areas.

2.3.2.2 Neural correlates of expectation violation (‘prediction error’)

There are many experimental paradigms that can be used to test how prior knowledge or expectations shape perceptual responses, including manipulations of context, priming, oddball tasks, and probabilistic learning, yet responses to violations of expectation are largely robust across measures and tasks.

First, fMRI has shown that unexpected visual targets are associated with an increase in BOLD amplitude over stimulus-specific sensory cortices (den Ouden, Friston, Daw, McIntosh, & Stephan, 2009; Eickhoff et al., 2011; Iglesias et al., 2013; John-saaltink, Utzerath, Kok, & Lau, 2015; Kok et al., 2011) that cannot be explained as repetition suppression (Larsson & Smith, 2012; Summerfield, Monti, Trittschuch, Mesulam, & Egner, 2009). For example, unexpected presentations of face stimuli are associated with an increased BOLD response in fusiform face areas (FFA) but not parahippocampal place area (PPA), whereas unexpected presentations of house stimuli are associated with increased BOLD in PPA but not FFA (Egner, Monti, & Summerfield, 2010). While these responses tend to occur in stimulus-specific regions, they can also be observed across cortex (Bubic, von Cramon, Jacobsen, Schröger, & Schubotz, 2009) and, remarkably, this pattern is seen in sensory cortices even when a target is unexpectedly *absent* (e.g. Todorovic, van Ede, Maris, & de Lange, 2011). Predictive processing or Bayesian brain approaches interpret these mismatch, or ‘surprise’ responses as reflecting perceptual ‘prediction error’.

We see the same expectation violation, or mismatch, response in ERP research. One particularly relevant example is the mismatch negativity (MMN), elicited in oddball tasks when rare, deviant tones are presented within standard auditory sequences. Because this evoked response reflects a mismatch between expectation (the more frequent tones) and sensation (the deviant), the MMN has been interpreted in predictive coding terms as a neural signature of prediction error (Stefanics et al., 2014; Wacongne, Changeux, & Dehaene, 2012). In a variant of the oddball task, individual tones can be local oddballs and standards, but also entire sequences can be global oddballs or standards, depending on whether they rarely or commonly occur. For example, suppose the experimenter plays two tones: A and B. If they are presented in the sequences AAAAB AAAAB AAAAB AAAAA, the final ‘A’ tone is a local standard because it is preceded by a series of 'A's, but it is embedded in the sequence AAAAA, which is a global deviant. Using this design, Chennu and colleagues (2013) lent (indirect) support for another prediction of hierarchical predictive coding: that mismatch responses at higher regions of the cortical hierarchy are associated with violations of more abstracted or complex rules.

Violations of expectation are associated with an increased ERP and BOLD response, but also an increase in high frequency, gamma-band oscillatory activity. Gamma oscillations have long been associated with stimulus-driven processing (Börgers & Kopell, 2008; Buzsáki & Wang, 2012; Donner & Siegel, 2011; Kopell, Kramer, Malerba, & Whittington, 2010), and predictive processing frameworks hypothesise this frequency channel to be the mechanism by which prediction errors are fed forward up the cortical hierarchy (Bastos et al., 2012). Consistent with this view, evoked gamma-band oscillations are predominately communicated via feedback connections (Bastos et al., 2015; van Kerkoerle et al., 2014; von Stein, Chiang, & König, 2000). Furthermore, convergent evidence from both model-based analyses and experimental manipulations of prior probability have revealed that increased evoked gamma power is positively associated with the violation of expectations (Arnal, Wyart, & Giraud, 2011; Brodski, Paasch, Helbling, & Wibral, 2015; Brunet et al., 2014; Pelt et al., 2016; Summerfield & Mangels, 2006). Recently, the association between gamma oscillations and prediction error has received support from model-based electrocorticography, Sedley and colleagues show that local field potentials in the gamma band show a robust (albeit small) correlation with prediction error, but that gamma oscillations are better explained as precision-weighted prediction error (Sedley et al., 2016).

2.3.3 The relationship between top-down attention and prior expectation

The data reviewed so far suggests that both prior expectations and top-down attention shape perception. At the behavioural level, attention primarily seems to target sensitivity, while expectations primarily seem to bias judgements towards those that are more likely to be correct. There is also converging evidence from a range of techniques that suggests attention increases the neural response to task-relevant features, whereas expectations suppress responses to stimuli that have been predicted. This motivates the important question of whether and how they interact.

Bayesian decision theory proposes that priors should carry greater weight on decision under uncertainty (see Chapter 1), but this does not seem to hold empirically. At the level of behaviour, expectations decrease reaction times and bias responses independently of attention (Kok et al., 2012), though under inattention priors are incorporated into decision sub-optimally (Morales et al., 2015). Evidence from functional neuroimaging and EEG may also lie contrary to the hypothesised relationship between priors and attention. Some research shows that inattention decreases the effects of expectation. For example, perceptual ‘prediction error’ responses to unexpected faces and houses are attenuated or absent under inattention (Jiang, Summerfield, & Egner, 2013; Larsson & Smith, 2012, but see Kok et al., 2011), and attention is necessary for face and house priors to be decoded from category-specific areas (Jiang et al., 2013).

By contrast, some research that shows expectations shape neural responses independently of attention. For example, priors about stimulus orientation can be decoded from V1 independently of attention (Kok et al., 2012), and deviant tones are associated with increased amplitude of early components of visual evoked potentials (VEPs) independently of attention (Hsu, Hämäläinen, & Waszak, 2014b). While further research is needed to understand the relationship between attention and expectation, it may be that the interaction between attention and expectation is dependent on the stimuli. Surprising gratings and pure tones seem to induce mismatch responses independently of attention, whereas these responses seem to require attention when using face and house stimuli. Thus, it may be that the dependence of expectation on attention is associated with the extent to which stimulus-selective processing can be achieved pre-attentively: the detection of low-level features (for which there exist specialised neurons) can be achieved without attention, whereas more complex perceptual tasks, such object categorisation, cannot (VanRullen & Thorpe, 2001). Importantly, if this account can explain why some studies find a dependence of expectation and some do not, it cannot explain why results consistently deviate from Bayesian principles.

2.4 SUBJECTIVE PERCEPTUAL CONFIDENCE

Having seen that top-down attention and top-down expectation have dissociable effects on perception, sharpening versus biasing perceptual decisions, we can ask how subjective perception may be influenced by these processes. Subjective decision confidence is constructed from an introspective process, in which the subject estimates the probability that their decision was correct (see Chapter 3 for how this can be measured and formal treatments of decision and confidence). For example, I may be able to identify a person in front of me in daylight easily, but at night their features may be harder to detect, and I become unsure of whom I have encountered. In the former case, I will estimate my chance of having correctly identified the person at close to 100%, while in the latter it will likely reduce substantially in the absence of non-visual cues. Confidence (my estimate) differs from uncertainty (daytime or night-time), because while confidence is defined with respect to a decision, uncertainty is not (Pouget et al., 2016). Rather, uncertainty pertains to variability, and can be conceived as a distributional property of a state (the degree of ‘internal noise’ or the reliability of the representation), the stimulus (e.g. stimulus noise), the environment (volatility, that is, changes in the statistics of the world), or action-outcome mappings (Bach & Dolan, 2012).

We know that reports of subjective confidence reflect meaningful estimates of decision accuracy, because they strongly correlate with task performance (for a review, see Grimaldi, Lau, & Basso, 2015) and task difficulty (Baranski & Petrusic, 1998; Maniscalco & Lau, 2012; Vickers, 1979). However, it is not clear how we are able to estimate our decision accuracy well, especially in the absence of external feedback. How do we *know* what we see?

2.4.1 Perceptual decision and confidence are tightly related but dissociable

It is clear that confidence is constructed, in part, from the same evidence used for the objective decision. First, if a subject is probed for a response during evidence accumulation, confidence increases with time, reflecting increased accumulation of evidence (Baranski & Petrusic, 1998; Festinger, 1943; Vickers & Packer, 1982). However, for un-speeded responses, high confidence is associated with faster reaction speeds, possibly reflecting faster evidence accumulation (Vickers & Packer, 1982; Douglas Vickers, Smith, Burt, & Brown, 1985). We are also more confident in decisions when sensory uncertainty is low, or when the target is easier to discriminate (Baranski & Petrusic, 1998; Maniscalco & Lau, 2012; Peirce & Jastrow, 1884; Spence et al., 2015). Finally, confidence and decision evidence share common neural signatures at early stages of the decision-making stream, suggesting that they arise from the same information source (Fetsch, Kiani, Newsome, & Shadlen, 2014; Gherman & Philiastides, 2015; Kiani, Corthell, & Shadlen, 2014; Kiani & Shadlen, 2009).

While decision and confidence are tightly related, confidence must incorporate additional evidence. We know this because decision accuracy and metacognitive accuracy (the correspondence between confidence and accuracy) can dissociate. For example, we can have above chance performance, yet no confidence our decisions (Marcel, 1993; Overgaard & Sandberg, 2012), which is best illustrated in blindsight patients (Ko & Lau, 2012; Leopold, 2012). These patients can respond to objects in their visual field, yet are not conscious of those objects. We can also ‘know that we do not know’, that is, be at chance accuracy while constructing confidence judgments that are sensitive to that lack of evidence (Scott, Dienes, Barrett, Bor, & Seth, 2014). In this latter example, if a subject has insight to their lack of knowledge that is purely based upon objective decision evidence, they ought to have selected the alternative response.

These findings refute models that assume confidence to be a read-out of decision evidence (e.g. Ratcliff & Starns, 2013; Vickers, 1979). Adaptations of these models allow evidence to accumulate after the objective decision has been made (e.g. Pleskac & Busemeyer, 2010). In these models, dissociations between decision and metacognitive accuracy are easier to accommodate. Furthermore, they are supported by evidence for ERP signatures of the decision variable, the error positivity (Pe), that continue to evolve after erroneous decisions such that evidence for subjective uncertainty is accumulated (Murphy, Robertson, Harty, & O’Connell, 2015).

Although these models are promising, they are not particularly adept in accounting for asymmetries in how evidence is accumulated across objective and subjective decisions. For example, confidence is more sensitive to uncertainty than the objective judgement (Spence et al., 2015), and while uncertainty tends to be optimally weighted in objective judgements, it is underestimated in confidence judgements (Zylberberg et al., 2012). In a similar vein, nonconscious decision evidence can shape objective decisions but not subjective decisions (Vlassova, Donkin, & Pearson, 2014). Speculatively, it may be that uncertainty is incorporated into confidence judgements twice: first, sensory noise may primarily shape evidence accumulation, and second, confidence may additionally incorporate estimates of internal noise. However, other findings, such as the invariance of confidence judgements to evidence for un-chosen decisions remains problematic (Maniscalco et al., 2016; Zylberberg et al., 2012).

2.4.2 Neural correlates of confidence

In humans, prefrontal and frontal areas seem to play a particularly important role in confidence and metacognitive monitoring (Fleming & Dolan, 2012; Grimaldi et al., 2015; Yeung & Summerfield, 2012). In particular, dorsolateral prefrontal cortex (DLPFC, e.g. Lau & Passingham, 2006; Rounis, Maniscalco, Rothwell, Passingham, & Lau, 2010), rostrolateral prefrontal cortex (rlPFC, De Martino, Fleming, Garrett, & Dolan, 2013; Fleming, Huijgen, & Dolan, 2012; Fleming, Weil, Nagy, Dolan, & Rees, 2010), and orbitofrontal/ventromedial prefrontal cortex (OFC/vmPFC, e.g. De Martino et al., 2013; Lebreton, Abitbol, Daunizeau, & Pessiglione, 2015; Rolls et al., 2010b; Yokoyama et al., 2010) are frequently implicated in metacognitive tasks. The default mode network has also been implicated in decision confidence (White, Engen, Sørensen, Overgaard, & Shergill, 2014).

However, distinguishing confidence from its antecedents, such as the decision variable, modulators of confidence, such as expected reward, and ensuing processes, such as action plans, has proved difficult. Accordingly, many prefrontal regions implicated in confidence are also implicated in processes such as response preparation and evidence accumulation, particularly DLPFC (Mulder, van Maanen, & Forstmann, 2014). Fleming and colleagues isolated the process of introspecting on one's decision accuracy from motor plans, by comparing conditions in which participants reported their confidence and decisions in which participants gave the confidence report presented on-screen (Fleming et al., 2012). This manipulation implicated rostrolateral prefrontal cortex (rlPFC) in the construction or communication of confidence, a claim further supported by the findings that rlPFC grey matter volume predicts metacognitive accuracy (Fleming et al., 2010; McCurdy et al., 2013), and that patients with legions in this area exhibit domain-specific metacognitive impairments while leaving sensitivity unimpaired (Fleming, Ryu, Golfinos, & Blackmon, 2014). These results implicate rlPFC in a role that is distinguished from report and the objective choice. However what is unclear is whether confidence is 'computed in' rlPFC, that is, whether rlPFC integrates all relevant sources of information into confidence, or whether rlPFC exists within a post-decisional stream that iteratively refines the confidence estimate.

While processes that are distinct from objective decision-making must shape confidence, it is not clear what these processes are. One plausible candidate is error detection, potentially arising in anterior cingulate cortex and reflected in

ERP components ERN and Pe. Error detection may provide information on how likely the decision is to have been correct, given experienced response conflict or additional evidence accumulation (Boldt & Yeung, 2015; Charles, Van Opstal, Marti, & Dehaene, 2013; Steinhauser & Yeung, 2010). This is supported by recent evidence that the Pe component, represented in posterior medial frontal cortex, evolves after an erroneous decision has been made, reflecting post-decision evidence accumulation about decisional accuracy (Murphy et al., 2015). While confidence may be shaped by error detection systems, an explanation of confidence in terms of error detection alone does not explain how the brain is able to infer the accuracy of its decision, especially in tasks that do not induce high response conflict.

At least in part, this process appears to be domain-specific because metacognitive decisions in the memory and perceptual domain are subserved by distinct functional networks (Baird, Smallwood, Gorgolewski, & Margulies, 2013), where perceptual metacognition primarily recruits anterior prefrontal cortex, as well as anterior cingulate cortex, putamen, caudate and thalamus. This domain-specificity is consistent with evidence that patients with anterior prefrontal lesions show selective impairments in perceptual metacognition (Fleming et al., 2014).

Studies, especially those which implement model-based analyses, sometimes find representations of confidence in reward-related subcortical regions, especially striatum (Braunlich & Seger, 2016; Daniel & Pollmann, 2012; Hebart, Schriever, Donner, & Haynes, 2014), even in the absence of external feedback (Guggenmos et al., 2016). However activity in these regions might reflect implicit reward associated with believing a correct decision has been made, their association with surprise (Domenech & Dreher, 2010), task-difficulty (Green et al., 2013) or fluctuations in decision thresholds (Mansfield, Karayanidis, Jamadar, Heathcote, & Forstmann, 2011) that simply covary with changes in confidence.

2.4.2 Do top-down influences shape confidence, metacognition and awareness?

The key question addressed in this thesis asks whether top-down influences of expectation can shape subjective confidence. Little work has addressed whether top-down attention influences perceptual metacognition. Kanai and colleagues have shown that in attentional blink and dual-task paradigms, where attention is diverted either by the rapid presentation of distractors (attentional blink) or by a concurrent task (dual-task), participants are still able to detect attentional failures of awareness (Kanai et al., 2010). However when giving prospective confidence ratings, participants overestimate the cost of diverting attention from the primary task (Finley, Benjamin, & Mccarley, 2014).

The role of spatial attention in confidence judgements is also unclear. Work has both shown that confidence is invariant to spatial attention (Wilimzig & Fahle, 2008) and increases with spatial attention (Zizlsperger, Sauvigny, & Haarmeier, 2012). Surprisingly, Rahnev and colleagues found that subjective visibility ratings disproportionately *decrease* with attention (Rahnev et al., 2011). However, their data also revealed an interesting relationship between attention and subjective visibility and stimulus contrast: with attention, subjective visibility increases with stimulus contrast, as would be expected. However, visibility ratings under inattention are relatively invariant to stimulus contrast. This suggests that under inattention, the integration of sensory uncertainty into subjective judgements is impaired. The authors capture this in a signal detection model, where inattention leads to more variable perceptual representations (Gaussians with a larger variance). They show that using the same confidence thresholds for attended and unattended targets results in fewer attended than unattended targets being reported as highly visible. These empirical results are difficult to interpret from predictive processing perspectives, as invariance of judgements to uncertainty cannot be accommodated by changes in precision-weighting; rather, it suggests that prediction error arising from these unattended targets receives no precision-weighting at all. Alternatively, it may be that only goal-relevant targets are subject to context-sensitive gain control.

Indirect evidence from studies on visual awareness suggests that expected targets would be associated with higher confidence. This follows from work showing that expected targets receive preferential access to awareness (Chang, Kanai, & Seth, 2015; Pinto, van Gaal, de Lange, Lamme, & Seth, 2015). We also know that prior exposure to a stimulus increases both behavioural reports of subjective visibility, and neural correlates of subjective visibility such as ERP component N2 (Melloni et al., 2011). Prior experience with a stimulus also seems to have a stronger effect on subjective than objective decisions, because training participants on a perceptual task in one spatial location leads to increased subjective visibility at untrained locations, while leaving sensitivity unaffected (Schwiedrzik, Singer, & Melloni, 2011).

Interestingly, trial-by-trial confidence judgements seem to be shaped by expected confidence as well. Rahnev and colleagues showed that independently of spatial attention and stimulus contrast, confidence on one trial is positively associated with confidence on previous trials, but decision accuracy is not. Moreover, this ‘confidence leak’ persists when the task on previous trials differed from that on the current trial (Rahnev, Koizumi, Mccurdy, Esposito, & Lau, 2015). Consistent with this, Guggenmos and colleagues showed that confidence on a trial can be predicted from confidence on previous trials, and that the discrepancy between current and recent confidence may be associated with left ventral striatum (Guggenmos et al., 2016). Together, these results show that confidence cannot simply be a product of sensory processes, but must incorporate additional, confidence-specific influences.

2.5 Outstanding questions

This chapter has shown that top-down attention and top-down priors robustly shape objective perception, and that evidence for the perceptual choice is integrated into subjective confidence judgements. However, the relationship between top-down attention and expectation on confidence remains understudied, and the extant literature suggests that subjective judgements may not be affected by attention and expectation in the same way as objective judgements. The neural mechanisms underlying top-down influences on subjective judgements also remain largely unknown. This thesis investigates whether and how top-down priors and top-down attention shape metacognitive confidence judgements, orthogonally manipulating these two key processes in order to separate their respective effects.

In order to examine how perceptual priors influence confidence, this thesis implements a novel dual-task paradigm. Here, the critical task involves detecting a faint, peripheral Gabor target, and each block of trials is associated with a different prior probability of its presentation. Participants are informed of this prior probability before trials begin, however because the expectancy cues are valid ('true') they will be corroborated by the visual information sampled on each trial.

This manipulation of priors is contextual, in the sense that they do not pertain to within-trial probabilities (for example cue-target associations, e.g. Kok et al., 2011), but across-trial probabilities. This is important because trial-wise probabilistic information increases the risk of a trivial confidence attribution, especially for explicit cues, such that the participant may consciously derive their confidence report from the prior information.

Recent work has begun to delineate between effects of prior expectations and of attention on perceptual decision-making, and for this reason, all empirical chapters here manipulated attention and expectation orthogonally, while keeping detection sensitivity constant across conditions. The attentional manipulation used throughout the thesis was dual-task, where attention was either fully allocated to the critical target detection task, or shared with a central visual search task. Motivational salience associated with expecting Gabor presence should be minimised when attention is diverted. In Chapters 4 and 6, the Gabor target has a gradual onset and offset, minimising the chances of attentional capture.

3

INFERENCES ABOUT CONSCIOUSNESS USING SUBJECTIVE REPORTS OF CONFIDENCE

3.1 Overview

An important aspect of consciousness is the ability to reflect upon one's own thoughts, an insight which can be traced back to John Locke, who stated that "*consciousness is the perception of what passes in a man's own mind*" (Locke, 1700)*.* This definition of consciousness forms the basis of Higher Order Thought (HOT) theories of phenomenal consciousness (Gennaro 2004; Lau & Rosenthal 2011; Rosenthal 1986), which posit that it is those states for which we have some representation or conceptualisation that we have phenomenology for. It is not necessary to subscribe to this account of consciousness to appreciate that our ability to reflect upon our own thoughts and decisions taps into an important facet of awareness. We can operationalise the ability to evaluate these decisions as metacognitive sensitivity or metacognitive accuracy, terms used interchangeably here. These are defined as the ability to judge the correctness of one's own decisions. We say that metacognitive accuracy is high when decision confidence exhibits a positive association with decision accuracy. So, a subject with high metacognitive accuracy 'knows when they know', and will largely ascribe high confidence to correct decisions and low confidence to incorrect decisions. This Chapter will discuss ways in which confidence reports can be collected and ways in which confidence and metacognition can be measured.

First, this Chapter will present a brief overview of type 1 Signal Detection Theory (SDT), used throughout this thesis to characterise objective task performance and decision biases. A more thorough account is given in Macmillan & Creelman (2004) and Green & Sweets (1966). This Chapter will also cover ways in which the researcher may want to measure confidence, and what we ultimately need from a good metacognitive measure. Next, it will move to a discussion of measures of metacognition and confidence. These quantify metacognitive accuracy by examining the correspondence between decision accuracy and decision confidence. Specifically, it will first cover correlation measures Pearson's *r* and the phi correlation coefficient, and then move to measures type 2 *D'*, type 2 ROC curves, and finally *meta-d'*.

3.2 Characterising objective performance

3.2.1 Type 1 SDT

Signal detection theory (SDT. Green & Swets 1966; Macmillan & Creelman 2004) models the way in which we make binary choice perceptual decisions. Here, the participant has to choose whether they should attribute stimulation to just noise, or to a noisy signal. Alternatively, the model can capture the choice between a noisy 'type A' signal and a noisy 'type B' signal. In this Chapter, we will consider the 'absent' versus 'present' scenario. However, all the methods work equally well for `A’ versus `B’: “yes” can simply be replaced by “left orientation” or “old word”, and “no” can simply be replaced by “right orientation”, “new word”, and so on.

The signal detection model is illustrated in figure 3.1. SDT assumes that we can represent the probability of some decision evidence having been caused by target absence as a Gaussian probability density function (depicted in red). We can do the same for the target present case (depicted in blue). This evidence, represented on the x-axis, corresponds to an internal state induced by the signal. In this way, stronger internal representations of the signal, for example of stimulus contrast, will be associated with a higher probability of target presence and a lower probability of target absence. SDT assumes that the evidence required to report either option is determined by the decision threshold, which bifurcates the decision axis into evidence that will result in a “yes” versus a “no” response.

This decision threshold, or criterion, is modelled as a horizontal intercept called c or θ. An unbiased observer will set their decision threshold at the intersection of the ‘target absent’ and ‘target present’ Gaussians, as is the case in figure 3.1. If the decision axis is aligned so that the peak of the ‘target absent’ Gaussian is at zero and has a standard deviation of 1 we will get $c = 0.5$. In this Chapter we will place the distributions such that an unbiased criterion gives $c = 0$.

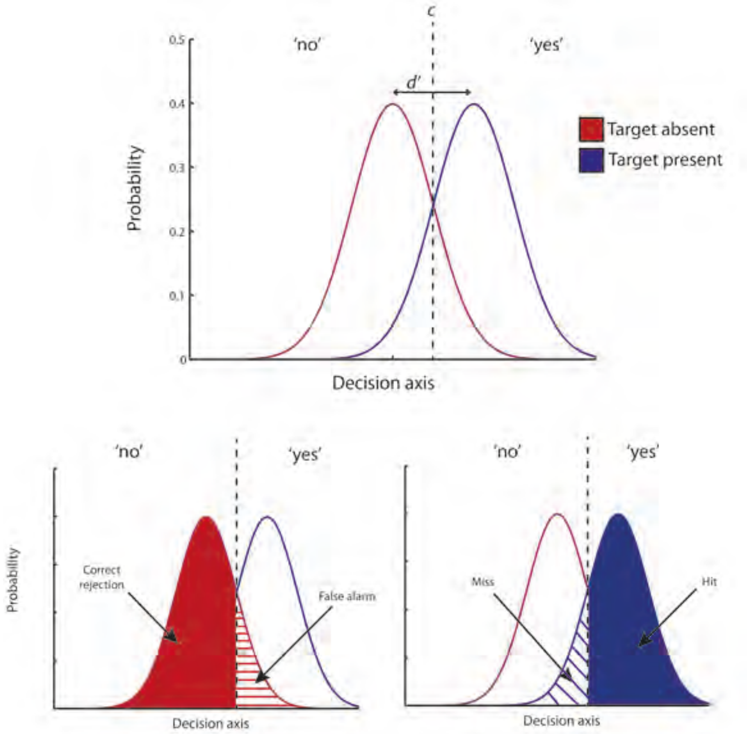

Figure 3.1. Type 1 signal detection theory.

Top panel. Probability of the internal representation having been caused by target absence (red Gaussian) or target presence (blue Gaussian). Sensitivity *d'* is defined as the separation between the peak of the two Gaussians. Decision threshold *c* is represented as a black dashed line. If decision evidence exceeds the threshold then the participant will make report "yes", and if it is less then they will report "no". Here *c* is equally placed between the two Gaussians and therefore responses are unbiased.
Bottom panel. These figures illustrate how we can predict choices from this model. On the bottom left the target is absent. 'No' responses in this case are correct rejections, whereas 'yes' responses are false alarms. On the bottom right panel the target is present. Here, reporting 'yes' is a hit and reporting 'no' is a miss.

Table 3.1 SDT responses

	Respond "present"	*Respond "absent"*
Signal present	*Hit*	*Miss*
Signal absent	*False Alarm*	*Correct Rejection*

In this way, positive values of *c* are conservative, corresponding to a bias towards reporting “no”. Negative values are liberal, corresponding to a bias towards reporting “yes”.

Detection sensitivity *d’* is defined as the difference between the peaks of the two Gaussians, and is given in units of the standard deviation of the target absent distribution. Higher values of *d’*, indicated by a greater separation between the two distributions, correspond to higher sensitivity because there is a smaller portion of decision evidence that can support both choices. We can also compute relative *c,* denoted *c’* and defined as *c* /*d’*. Here, *c* is taken relative to the distance between the two Gaussians. This measure quantifies how extreme the criterion is, relative to performance.

If the assumptions of SDT are met, sensitivity *d'* will be invariant to decision bias *c*. The first assumption is that the two probability density functions are indeed Gaussian. In low-level perceptual tasks this assumption holds, because by the central limit theorem, the distribution of activity of a large body of sensory neurons responding to the target will approach normality. The second assumption is that the two Gaussians have equal variances. It is this second assumption that tends to be problematic in psychology research; however if an unequal variances model fits better, then the corrected d'_a (see below) can be used instead. For example, yes/no tasks are thought of as being fit best by an unequal variances model.

In order to calculate *d’* the researcher collects data in a 2 x 2 design such that a signal is present or absent and the participant can be correct or incorrect. This leads to a table of response variables as shown in Table 3.1 and the bottom panel of figure 3.1. We can then calculate the following:

$$Hit\ rate = \frac{\sum Hits}{\sum(Hits + Misses)}$$

$$False\ alarm\ rate = \frac{\sum False\ alarms}{\sum(False\ alarms + Correct\ rejections)}$$

From these, task performance *d'* and decision threshold *c* can then be calculated as

$$d' = \phi^{-1}(hit\,rate) \;-\; \phi^{-1}(false\;alarm\,rate)$$

$$c = \frac{-\phi^{-1}(hit\,rate) - \phi^{-1}(false\;alarm\,rate)}{2}\;,$$

where ϕ^{-1} is the inverse cumulative probability density function of the standard Gaussian distribution (also commonly known as the Z-statistic). These statistics are in the units of the standard deviation of the noise distribution when its mean is set to zero.

If the researcher is assuming an unequal variance model, adjusted *d'$_a$* can be calculated as

$$d'_a = s\phi^{-1}(hit\,rate) - \phi^{-1}(false\;alarm\,rate),$$

where *s* is the ratio of the standard deviation of the signal-plus-noise distribution to that of the noise distribution.

By collecting confidence ratings, we can estimate *s* from data. From these, the researcher can obtain hit and false alarm rates for multiple decision thresholds (as described below in the section 3.1). Subsequently, *s* and *d'$_a$* can be computed from the best-fit values for the above equation for all hit rates and false alarm rates*.* However, if *s* has not been estimated to a good degree of accuracy we cannot assume that *d'$_a$* and *c* are (approximately) independent (Macmillan & Creelman 2004). Furthermore, it may be problematic to infer *s* from confidence ratings that are subsequently used for further SDT analyses on confidence (Maniscalco & Lau, 2014).

3.2.2 TRANSFORMING DATA WITH ZERO OR UNITY HIT RATE OR FALSE ALARM RATE.

There are occasions when one obtains hit rates or false alarm rates of zero or one. In these cases, data have to be transformed to avoid infinities in the equation for *d'*. These arise from the ϕ^{-1} function going to plus/minus infinity at

1/0. For *d'* to be finite, the hit and false alarm rates always lie strictly between 0 and 1. For this reason, extreme hit and false alarm rates, that is, hit rate > 0.95 or false alarm rate < 0.05, may generate unstable estimates of *d'* and *c'.*

In most cases, these situations can be avoided by ensuring that one collects a large number of trials per condition (at least 50) and that manipulations that may affect the decision criterion, for example performance-related reward or punishment, are not too strong. However, such a manipulation may be the focus of the experiment (as is the case in this thesis). In this case, extreme data are obtained then the researcher can use one of two main transformations. In one, the researcher only adapts problematic data. Here, in an experimental set-up with *n* signal present trials and (*N* - *n*) signal absent trials, a zero hit or false alarm rate would be replaced with 1/2*n* or 1/2(*N* - *n*) respectively. A hit or false alarm rate equal to one would be replaced with 1-(1/2*n*) or 1-(1/2(*N* - *n*)) respectively. Thus, each of these variables is transformed proportionately to the number of trials across which it is computed. For example, in the case that 25% of 100 trials are signal trials, a 0 or 1 hit rate would be shifted by 1/50 and a 0 or 1 false alarm rate by 1/150. This method is called the 1/2N rule (Macmillan & Kaplan 1985).

An alternative transformation, the log-linear transformation, was proposed by Snodgrass & Corwin (1988). Here, all data cells (total hits, false alarms, correct rejections and misses), regardless of whether they are problematic or not, have 0.5 added to them. This is advantageous in that all data are treated equally, and in that it impossible to have zero or one hit or false alarm rates. This correction can be considered a (Bayesian) prior on a *d'* and *c* of zero (Barrett, Dienes, & Seth 2013; Mealor & Dienes 2013).

Hautus (1995) modelled the effects of both of these transformations on *d'* and found that both transformations can bias *d'* measures. While the log-linear rule systematically underestimated *d'*, the 1/2N rule was more biased, and could distort *d'* in either direction. Therefore, although the log-linear rule is recommended over its counterpart, minimising the risk of collecting data with empty cells is preferable.

3.2.3 Type 1 ROC curves

If the assumptions of SDT have been violated we can create a model-free receiver operating characteristic (ROC) curve, which plots hit rate against false alarm rate for possible type 1 thresholds. Accordingly, the area under the ROC (AUC or AROC) gives us a measure of detection sensitivity that is also invariant to response biases. Plotting the ROC curve requires participants to select a stimulus class (S_1, versus S_2), for example from 1 = definitely S_1 to 6 = definitely S_2. Then, the researcher plots hit rate against false alarm rate via hypothetical decision criteria based on different thresholds of the responses. If a response scale of length *n* has been used then there are *n* - 1 ways to partition responses into hypothetical levels of decision criterion. Each partition determines the boundary between S_1 and $S_{2.}$ For example, first we would partition the data such that a rating of 1 indicates an S_1 response and a rating of 2-6 indicates $S_{2.}$

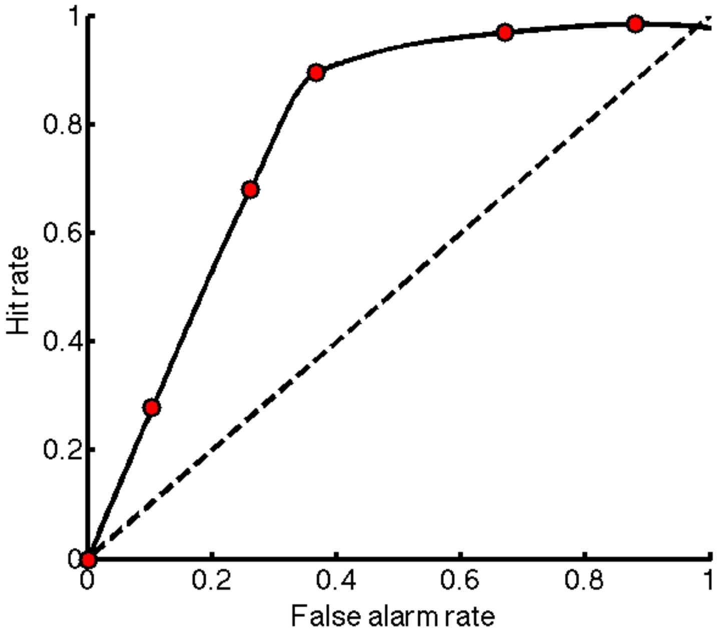

Figure 3.2. Type 1 ROC curve.

Each red circle corresponds to a (Hit rate, False alarm rate) pair, derived from a different partition of the response scale. These pairs shown us how the relationship between hit rate and false alarm rate change as decision threshold varies. The black dashed line indicates chance performance. The area under the curve (AUC) is the area between the chance line and the ROC function.

Then, one would partition such that a rating of 1 or 2 indicates an S_1 response and 3 to 6 indicates S_2, continuing until a rating of 1-5 indicates an S_1 response and a rating of 6 indicates an S_2 response. Therefore, for each hypothetical decision threshold one obtains different numbers of hits and false alarms. From these, hit and false alarm rates can be computed. As shown in figure 3.2, these are plotted against each other, producing a curve that characterises sensitivity across a range of decision biases without making assumptions about the underlying signal and noise distributions. The diagonal on the graph represents chance performance. The more than the ROC curve extends above the diagonal, the greater the sensitivity, in that for any given false alarm rate the corresponding hit rate is higher. Thus, the area under the ROC curve represents task performance.

It should be noted that because it does not rely on the assumptions of SDT, ROC curve analysis is not technically SDT. If one does assume that decisions are made based on an SDT model (without necessarily assuming an equal variance model), then the Z-transform of the ROC curve can be taken. This is a straight line, and the area under the (non-transformed) ROC curve can be obtained from a simple formula in terms of the slope and intercept of the Z-transform:

$$A_z = \Phi\left[\frac{Intercept}{\sqrt{1 + slope^2}}\right]$$

One potential problem with estimating type 1 ROC curves from confidence ratings is that they may conflate type 1 and 2 performance (Galvin, Podd, Drga, & Whitmore, 2003; Maniscalco & Lau, 2014). Additionally, if the researcher wants to examine changes in decision bias this will not be an appropriate analysis. However, a benefit of plotting an ROC curve or using SDT's *d'*, is that task performance can be decomposed into possible drivers of the change: hit rate and false alarm rate. For example, some empirical questions might hypothesise a change in hit rate but not false alarm rate. Kanai, Muggleton & Walsh (2008) found that transcranial magnetic stimulation (TMS) over intraparietal sulcus induces perceptual fading by demonstrating such an asymmetry: Although *d’* reduced with TMS, this was driven by a decrease in hit rate only. If false alarm rate (reduced sensitivity on target absent trials) had also increased with TMS then the *d’* effect could not have been driven by perceptual fading, which by definition only affects target present trials. Rather, a concurrent reduction in false alarm rate would have implicated intraparietal sulcus in general perceptual performance.

3.3 Measuring metacognition: precursors

3.3.1 Measuring metacognitive accuracy

In order to investigate metacognitive judgements the researcher needs to collect both an objective judgement and a subjective judgement. Typically, experimental designs include some objective task, such as target detection or word recall, in which objective performance can be measured. To measure metacognitive judgements we use what is known as a ‘type 2 task’, a term first coined by Clarke, Birdsall & Tanner (1959) and Pollack (1959), and so-called in reference to the aforementioned type 1 task of making decisions or judgements about the ‘state of the world’. The type 2 task requires the participant to evaluate the accuracy of their decision. Galvin, Podd and Whitmore (2003) discuss the type 2 task and argue that

> “*…The fact that the second decision [confidence that the trial was a signal trial] is a rating and follows a binary type 1 decision does not make it a type 2 decision. If the second decision is a rating of confidence in the signal event rather than in the correctness of the first decision then it is a type 1 rating, no matter when it occurs.*”

Following this, it is advised that the confidence judgement requested refer to the accuracy in the participant’s decision. However from the perspective of consciousness science it seems counterintuitive to assume a distinction between asking for confidence in the signal and asking for confidence in the participant’s judgement; this suggests an asymmetry in the trustworthiness of the objective (type 1) and subjective (type 2) responses. If we instead assume type 1 decisions refer to the state of the world (e.g. target presence versus absence), we can take type 2 decisions as probing the mental state or representation the subject has of the target. In this sense the prompt “Confidence” should be equivalent to the prompt “Confidence that you are correct”, though this has not been addressed empirically.

3.3.2 Collecting confidence ratings

The traditional method of collecting confidence ratings is in two-steps: the judgement is made and then confidence is given, either in a binary fashion or on a scale. Whether confidence is collected on a scale or in a binary fashion will dictate the metacognitive measures available to use. Confidence scales (e.g. from 1 to 4) have the advantage of being more sensitive and they can later be collapsed into a binary scale, reducing the chance of getting 0 or 100% confident responses. However importantly, if conclusions about consciousness are to be drawn, we can only infer unconscious knowledge or unconscious perception from those trials where participants have reported no confidence. We cannot infer this from low confidence. Therefore, a rating scale should only be symmetrically collapsed into a binary scale if no conclusions are to be drawn about awareness.

If the question of interest relates only to perceptual awareness and does not wish to use SDT methods, the Perceptual Awareness Scale is an alternative method of collecting subjective reports. This scale asks participants to rate the subjective visibility of their percept on a scale of one (no perceptual content) to four (clear perceptual content). Because the conscious content itself is not probed, the scale is more effective for simple (e.g. detection) than complex (e.g. discrimination) designs (Dienes & Seth 2010; Sandberg et al. 2010).

These methods of collecting subjective ratings thus far have been ‘two-step’ procedures, which first request a type 1 report (e.g. yes/no) and then a type 2 report (e.g. confident/guess). An alternative method is to use a one-step procedure, whereby participants are asked to choose between two responses S_1 and S_2 and high and low confidence at the same time. For example, a rating scale could be used where the lowest value indicates high confidence in S_1 and the highest value indicates high confidence in S_2. This has the benefit of being a faster reporting procedure. Furthermore, because both decisions are reported at the same time this method may be preferable if the researcher wishes to minimise the difference in decision evidence available for the type 1 and 2 reports. Indeed, in the perceptual domain, reaction times between one-step and two-step procedures differ while leaving the confidence-accuracy correlation unchanged (Wilimzig & Fahle 2008).

3.3.3 WHAT MAKES A GOOD MEASURE OF METACOGNITION?

In order to assess the ability of an individual to monitor the accuracy of their decisions we need to be able to separate the information on which their decision is based from the insight into that information that they hold. Accordingly, Maniscalco and Lau distinguish between absolute and relative metacognitive sensitivity (Maniscalco & Lau, 2012). Absolute metacognitive sensitivity refers only to the relationship between confidence and accuracy, whereas relative sensitivity refers to the efficacy of the metacognitive evaluation, with no confound of information quality.

To illustrate, suppose that a participant shows higher metacognitive accuracy in task A than task B. It may be that this occurs because the participant has more decision evidence in task A. In this first scenario, both *d’* and metacognition will change. Alternatively, metacognition may have increased over and above any changes in *d’.* In the former example *absolute* metacognitive accuracy has changed, whereas in the latter *relative* metacognitive accuracy has changed. If one wants to measure relative rather than absolute metacognitive sensitivity, objective performance should be equated across conditions. This is also important if the researcher wants to measure subjective confidence, because confidence will generally increase with increasing task performance (Grimaldi et al., 2015). Similarly, if the researcher is interested in metacognitive *bias,* that is, the tendency of the participant to report with high confidence, evidence should also stay constant across conditions and participants.

Many measures of metacognition are biased by objective performance, decision thresholds and/or metacognitive bias (discussed later in this Chapter). We therefore need a metacognition measure that is invariant to these factors, or at least allows us to separate them. For example, by demonstrating reduced perceptual metacognition after theta-burst transcranial magnetic stimulation to prefrontal cortex, Rounis et al. (2010) were able to implicate this area in metacognitive sensitivity. They used bias-invariant (type 2) meta-*d’* (to be discussed later) as their measure, which allowed them to rule out the alternative interpretation that PFC is involved in determining confidence bias.

It is important to note that dependence on decisional or confidence biases is not problematic if one is aiming more simply to rate the subject’s performance on the type 2 task. Viewed this way, metacognition may be facilitated *because* of shifts in metacognitive bias. Signal detection theoretic methods are useful because they allow us to consider the above points. By enabling the calculation of response and confidence biases as well as type 1 and 2 performance one can see how measures of task performance and decision bias interact. Further, one can see whether improvements in metacognitive performance can be attributed (at least in part) to specific changes in behaviour, for example, increased confidence specifically for correct reports.

One also has to consider whether to obtain a single measure of metacognition across all trials, or whether to assess metacognition separately for each possible class of type 1 response, i.e. to use a so-called `response-conditional’ measure of metacognition. For example, in a target detection experiment, one has the classes “respond present” and “respond absent” (see table 3.3). Kanai, Walsh & Tseng (2010) defined the Subjective Discriminability Index (SDI) as a measure of subjective unawareness of stimuli, based on response-conditional metacognitive measures. Specifically, by using only trials where subjects reported absence of a target (type 1 correct rejections and misses) in the type 2 calculation, they obtained a measure of metacognition for perception of absence. Their logic was that chance metacognitive accuracy implies blindness to the stimulus, whereas above chance metacognitive accuracy implies that, although the subject reported the target as unseen, some perceptual awareness must have been present (inattentional blindness). This follows from participants' ability to modulate their post-decisional confidence according to their accuracy.

3.4 Confidence-Accuracy correlations

The most intuitive measure of metacognition would tell us whether accuracy and confidence are significantly, and highly, correlated. Two main alternatives are available: Pearson's *r* and *phi*. These are equal in the binary case, but distinct for the non-binary case (that is, if confidence is reported on a scale, the former can be used).

For paired variables *X* and *Y* corresponding to confidence and accuracy values for n participants, the correlation *r* between confidence and accuracy is calculated as

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{s_x}\right)\left(\frac{Y_i-\bar{Y}}{s_y}\right)$$

where s_x and s_y are the sample standard deviations of *X* and *Y* respectively.

Alternatively, the *phi* correlation coefficient is calculated as

$$\varphi^2 = \frac{\chi^2}{n}$$

where χ^2 is the chi-squared statistic and n is the number of participants.

When *X* and *Y* are binary, e.g. *X* equals 0 for low confidence and 1 for high confidence, and *Y* equals 0 for incorrect and 1 for correct, *phi* and *r* are equal to each other, and can be calculated from the formula

$$\varphi = \frac{n_{1,1}n_{0,0}- n_{1,0}n_{0,1}}{\sqrt{n_{.,1}n_{.,0}n_{1,.}n_{0,.}}}$$

where $n_{x,y}$ is the total number of trials on which $X = x$ *and* $Y = y$, and $n_{.,y}$ and $n_{x,.}$ are respectively the total number of trials for which $Y = y$ and $X = x$. Though simple, the problem with such a measure (and indeed, with any non-signal detection theoretic measure) is that *r* and ϕ can be inflated by bias without there being a true improvement in metacognitive accuracy. To illustrate, imagine a stimulus detection paradigm in which all participants perform at 70% accuracy. If one participant has a bias towards being confident whereas another tends to say they are guessing, the first of these participants will have a higher correlation between confidence and accuracy than the second without necessarily having increased insight into their own decision accuracy.

3.5 Goodman-Kruskal Gamma Coefficient

The Goodman-Kruskal Gamma coefficient, G (Goodman & Kruskal 1954) is a non-parametric analogue of the signal detection theoretic measure *d*'. Its appeal lies in its straightforward probabilistic operationalization, which overcomes problems surrounding assumptions about equal variance or normality. In its original form it is computed via the same 2 x 2 factors as *d'* and it can be extended to cases in which ratings are given on a response plus confidence scale (e.g. 1 = very confident no, 6 = very confident yes). By being distribution-free it hoped to also be a flexible measure of metacognitive accuracy when applied to type 2 data (Nelson 1984). Task performance V is characterised as follows for a 2 x 2 design, the construction aimed at eliminating dependence on overall response bias. Suppose there are two trials and one of them is `stimulus present' and one of them is `stimulus absent', and the subject responds `present' on one trial and `absent' on the other. Then V is the probability that these responses match the stimulus. The estimate of this (obtained from the data from all trials) is given by:

$$V = \frac{\Sigma\, hits \times \Sigma\, correct\ rejections}{(\Sigma\, hits \times\ \Sigma\, correct\ rejections) + (\Sigma\, misses \times \Sigma\, false\ alarms)}$$

The Gamma coefficient is then given by

$$G = 2V - 1 = \frac{(\Sigma\, hits \times \Sigma\, correct\ rejections) - (\Sigma\, misses\ \times \Sigma\, false\ alarms)}{(\Sigma\, hits \times\ \Sigma\, correct\ rejections) + (\Sigma\, misses\ \times \Sigma\, false\ alarms)}$$

To assess metacognitive performance, pairs of responses (on the confidence scale) are combined to produce an analogue of V. There is no simple formula for the general (non 2 x 2) case, so for a thorough explanation we refer the reader to Masson & Rotello (Masson & Rotello, 2009).

In order to verify G's supposed invariance to bias and distributional assumptions, Masson & Rotello (2009) simulated datasets in which metacognitive sensitivity was fixed and calculated G. More specifically, a 2AFC task was modelled as two probability distributions representing each choice. The difference between the means of these distributions was adjusted on simulation runs such that "population gamma", calculated by randomly sampling from the distributions in order to approximate the proportion of cases where A>B, was fixed. It was then compared to the Gamma obtained when considering decision biases. Indeed, they found that G does get distorted by decisional biases. Moreover, this distortion increased when data were simulated from an unequal variance model, suggesting that the invariance under reasonable changes to distributional assumptions may not hold.

3.6 TYPE 2 D'

Type 2 signal detection theory extends the logic of its type 1 counterpart by using confidence reports to map onto detection accuracy (Kunimoto et al. 2001; Macmillan & Creelman 2004). It assumes that correct and incorrect responses can be plotted on a 'type 2' decision axis as Gaussian random variables, analogously to the signal and noise distributions in type 1 SDT. Under the assumption that confidence is based on the same evidence as the type 1 decision, the type 2 axis is a transformation of the type 1 axis. The distance between the peaks of the distributions gives us our measure of metacognitive sensitivity, type 2 *D'*.

As shown in figure 3.2, type 2 variables are computed analogously to type 1 variables, but instead of examining the correspondence between signal and response, response accuracy and confidence are compared. We define a type 2 hit as a confident and correct response, a type 2 false alarm as a confident but incorrect response, a type 2 miss as a correct but unconfident response and a type 2 correct rejection as an appropriately unconfident, incorrect response.

Metacognitive accuracy measure type 2 *D’* is then calculated analogously to type 1 *d’*. We subtract the normalised type 2 hit rate from the normalised type 2 false alarm rate. The type 2 criterion, denoted *C*, represents metacognitive bias. This quantity reflects the extent to which the subject is over or under-confident. We may wish to quantify metacognitive accuracy or metacognitive bias separately for “yes” and “no” decisions (as is done in Chapters 4, 5 and 6). To do this trials are further be separated according to whether the participant’s type 1 decision was “yes” or “no”.

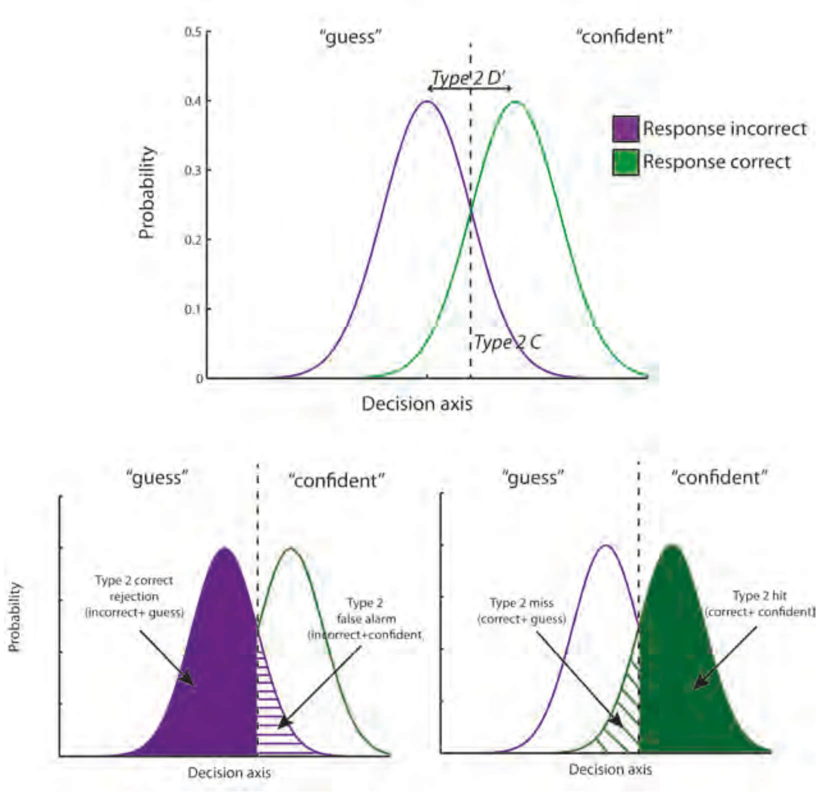

Figure 3.2. Type 2 D’.

These figures depict the probability of the response having been correct (green) or incorrect (purple). The separation between the two Gaussians gives us metacognitive sensitivity measure type 2 *D’*. The decision threshold type 2 *C* is placed somewhere on the decision axis. If the evidence strength is greater than this value the subject will respond ‘confident’, else the subject will respond ‘guess’. On the bottom left the decision was incorrect. ‘Guess’ responses in this case are type 2 correct rejections, whereas ‘confident’ responses are type 2 false alarms. On the bottom right panel the response was correct. Here, reporting ‘confident’ is a type 2 hit and reporting ‘guess’ is a type 2 miss.

Then, we can calculate separate measures of type 2 *C* and *D'* using separate sets of trials: “yes” trials and “no” trials (see table 3.2).

When the type 2 *D'* measure was proposed by Kunimoto et al. (2001) it generated much excitement, because it was thought to be invariant to bias like type 1 *d'.* Unfortunately, this is not the case. Kunimoto tested this measure with a wagering paradigm, where confidence was assessed by the magnitude of participants’ wager on each trail. Crucially, the total wager they could place was fixed for each session meaning that confidence biases were artificially fixed by the nature of the authors’ paradigm. Indeed, the claim that *D'* is invariant to metacognitive bias has been found to neither hold empirically (Evans & Azzopardi 2007) nor theoretically (Barrett et al 2013) when type 1 and 2 decisions are made based on the same evidence.

Table 3.2. Response-conditional type 2 response types

	Report “yes”		*Report “no”*	
	Correct	*Incorrect*	*Correct*	*Incorrect*
	Type 1 hit	*Type 1 FA*	*Type 1 CR*	*Type 1 miss*
Confident	*Type 2 Hit*	*Type 2 FA*	*Type 2 Hit*	*Type 2 FA*
Guess	*Type 2 Miss*	*Type 2 CR*	*Type 2 Miss*	*Type 2 CR*

FA false alarm, CR correct rejection

Barrett et al. (2013) found that under certain circumstances *D'* is highly unstable. For example, if the type 1 criterion is placed where the noise and signal and noise distributions intersect (i.e. *c* is unbiased) then *D'* is maximised when the observer is maximally unconfident, which would be a nonsensical and mal-adaptive strategy. By varying decision and confidence thresholds, Barrett and colleagues also found that *D*’ can range from being negative (which is difficult to interpret in a meaningful way) to being greater than type 1 *d*’. Importantly, these analyses demonstrate a high reliance of *D'* on decision and confidence thresholds. The behaviour of type 2 *D'*, then, does not suggest it to be a reliable measure of metacognition.

As well as being biased by decision and confidence thresholds, the validity of the underlying statistical assumptions of type 2 *D′* is also questionable. Specifically, the evidence for correct and incorrect responses cannot be represented as Gaussian distributions along a single decision axis (Maniscalco & Lau 2012). Galvin and colleagues have shown that no transformation of the type 1 decision axis will lead to the probability of correct and incorrect responses being normally distributed (Galvin et al., 2003). Therefore, if a single pathway underlies both type 1 and 2 decisions then *D'* is not a measure of metacognition that arises naturally from SDT modelling.

Despite these problems, in some scenarios type 2 *D'* can still be useful as a basic measure of type 2 performance, especially if response bias is small. For example, if there is an insufficient number of trials to use more robust measures then the researcher may wish to use *D′*. However, strong conclusions from analyses under type 2 D' and type 2 C should not be drawn.

3.7 Type 2 SDT

Type 2 *D'* is not the only way to envisage a type 2 signal detection theoretic model of metacognition. An alternative way to capture type 2 decisions is to overlay confidence thresholds onto the type 1 decision axis. As shown in figure 3.3A, confidence thresholds τ_+ and τ_- flank the type 1 decision threshold such that confidence is high when evidence X is less than τ_- or greater than τ_+ and low otherwise. Although this renders type 2 *D'* an unprincipled measure, it invites certain promising alternatives, namely type 2 ROC curves and *meta-d'*, as described in sections 3.5.2 and 3.5.3 respectively.

3.7.1 Confidence thresholds over the type 1 decision axis

As mentioned in section 3.4, type 2 *C* gives us a measure of confidence as a function of decision accuracy. However, the probability of making a correct response and the probability of making an incorrect response cannot be represented as Gaussians under any transformation of the type 1 SDT decision axis (Galvin et al., 2003). An alternative way in which we can model metacognitive decisions is to examine the type 1 evidence required for a decision to be reported with confidence. This entails placing two confidence thresholds onto the type 1 decision axis: one for “yes” responses, on the positive side of the type 1 decision threshold, and one for “no” responses which is placed on the negative side (Maniscalco & Lau, 2014). As these flanking thresholds approach the type 1 decision threshold θ (or *c*), that is, are tighter, less additional evidence is required for that decision to be reported with high confidence. So, tighter thresholds represent a more liberal metacognitive bias. This model can easily accommodate confidence scales with more than two points. One simply places more thresholds - one for each additional point on the scale. These must satisfy the assumption that the threshold for the lowest level of confidence extends the furthest out from θ, and increasing levels of confidence are associated with thresholds that are closer to θ.

The confidence threshold on the left of θ, τ-, tells us the type 1 evidence for absence needed to report ‘confident’, given that the subject has reported “no”. To estimate this we can reclassify trials on which the subject reported “no” with high confidence as simply “no” trials and all others as “yes” trials. Then, we recompute type 1 *d’* and *c* from these reclassified responses, which we call *d’*- and *c* - = τ-. Similarly, the threshold τ+ tells us the type 1 evidence needed for a “yes” response to be reported with confidence. This is computed analogously to τ-, but trials are reclassified as “yes” if the participant has reported target presence with confidence and “no” otherwise.

As shown in figure 3.3B and C, if the subject has responded “yes” or “no” the underlying probability distributions may drastically differ, which leads to differences in sensitivity for these two decisions. Indeed, we would expect the evidence for target presence to be higher if the subject responded “yes” than if they responded “no”. Because these distributions can differ, we cannot compare the response-specific confidence thresholds to θ. We need to normalise these thresholds by their respective response-specific *d*’s *d’+ and d’-.* We define the following:

$$C'_{1-} = \frac{\tau_-}{d'_-} \quad C'_{1+} = \frac{\tau_+}{d'_+}$$

where C'_{1-} is the normalised confidence threshold for negative responses over the type 1 axis, and where C'_{1+} is the normalised confidence threshold for positive responses over the type 1 axis. These values C'_{1-} and C'_{1+} tell us how extreme the confidence threshold is, in turn informing us of how over- or under-confident decisions are.

Now suppose the researcher wishes to determine the distance between each of C'_{1-} and C'_{1+} and θ (depicted in fig. 3.3A as distance A and B respectively). These distances are computed in empirical Chapters 4, 5 and 6 in order to measure effects of expectation on confidence after controlling for shifts in the type 1 decision threshold. For θ to be comparable to normalised values $C'_{1+}$and C'_{1-} we also need to normalise θ. Thus, we can take *c' = c / d'*, and compute the absolute difference between *c'* and each normalised confidence threshold. Taking the absolute difference is important because “no” decisions will have less “yes” evidence for than “yes” decisions, and accordingly always be placed to the left of θ. We do not want negative distances.

An important caveat here is that experimental manipulations that change C'_{1-} and C'_{1+} may be driven by changes in the variance of the type 1 distributions rather than by true metacognitive differences. This problem also arises when response-condition type 2 ROC curves are calculated (Maniscalco & Lau, 2014), and so this problem may similarly plague these confidence thresholds. Results using this measure should therefore be interpreted with caution.

3.7.2 Type 2 ROC Curves

While the type 1 ROC curve plots the probability of type 1 hits against the probability of type 1 false alarms for each level of criterion θ, the type 2 ROC curve plots the probability, for some fixed type 1 decision criterion *c*, type 2 hit rate against type 2 false alarm rate for all possible confidence thresholds. Just as type 1 ROC curves are robust to type 1 thresholds, so type 2 ROC curves should be robust to type 2 (and type 1) thresholds. However, because at the type 2 level there are two thresholds, τ_+ and τ_-, two parameters are left to vary freely (τ_+ and τ_-). This means that the type 2 ROC is not unique.

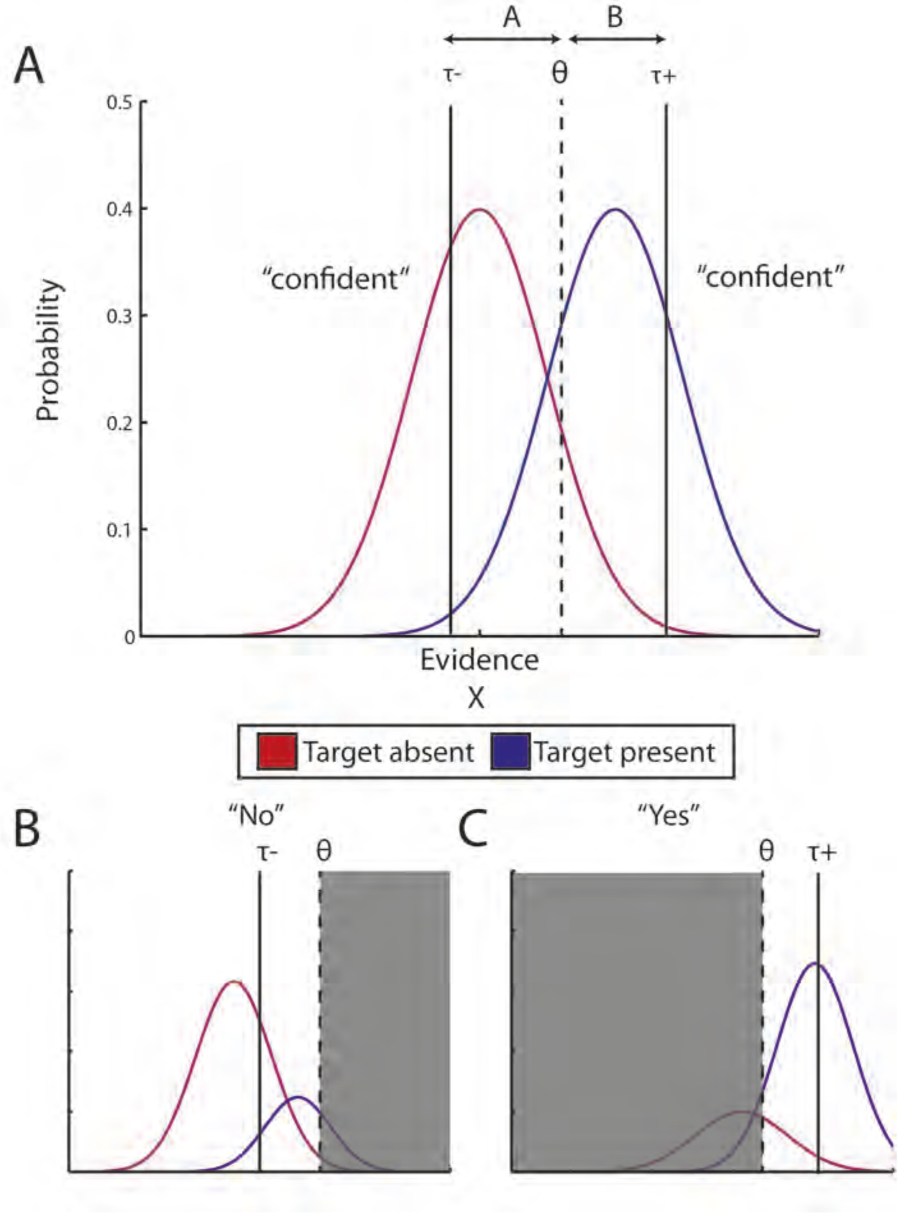

Figure 3.3 Type 2 signal detection theory

A. As opposed to the model illustrated in figure 3.2, flanking confidence thresholds are placed over the type 1 decision axis on either side of decision threshold θ. The confidence thresholds for "yes" and "no" responses are denoted τ_+ and τ- respectively.These need not be symmetric. As is the case for θ, the confidence thresholds indicate the type 1 evidence for the decision that is needed to report with high confidence. Given that the participant reports "no", the letter A between τ- and θ shows the additional evidence required for this judgement to be reported with confidence. The letter B shows the same but for "yes judgements. **B.** Here, the participant has reported "No". We see that the target present and absent distributions are very different to those in the top panel. As a result, even though the thresholds occupy the same points on the decision axis they will lead to different proportions of correct and incorrect responses. **C.** As for B, but for "Yes" responses.

The current literature posits three potential solutions to this. Galvin et al. (2003) suggested collapsing the two confidence thresholds into one likelihood function: the likelihood ratio of being correct versus incorrect. This enables a unique solution for fixed θ and is straightforward to compute. However, the authors still found a strong dependence of the area under the curve (AUC) on θ.

An alternative measure, proposed by Clifford et al. (2008) recommends comparing the type 1 ROC curve based on a confidence rating scale with the ROC curve obtained by manipulating *c* experimentally. That is, if we manipulate the physical properties of the stimulus such that response threshold changes (e.g. threshold contrast) then we can plot the false alarm rate/hit rate trade-off across artificially induced criterion shifts. This is the traditional type 1 ROC curve. We can compare this with an alternative type 1 ROC in which changes in criterion are modelled by differentially bisecting into "signal" and "noise", an n point rating scale all *n*-1 ways. If metacognition is SDT-optimal, these two ROCs should be the same. This point follows from the assumption that an optimal observer would fully use the same information for the type 1 and the type 2 decision. Thus, Clifford et al. proposed their divergence as a measure of metacognition. Again though, the degree of divergence is generally dependent on type 1 response bias.

Finally, Barrett et al. (2013) constructed the SDT-optimal type 2 ROC curve; the type 2 ROC curve that, for fixed θ and fixed type 2 false alarm rate (*F*), gives us the greatest type 2 hit rate (*H*), H_{max} (and therefore type 2 performance). Like the formulation above, this describes the performance of the SDT-optimal observer. Unfortunately this curve was also found to be vulnerable to distortions from θ however because it describes SDT-expected performance it can be used to check whether data conform to SDT. If the researcher wishes to measure metacognitive bias (confidence), type 2 ROC curves will not be appropriate because they attempt to eliminate bias by design.

3.7.3 Meta-d’ and Meta-d’-balance

Meta-d' measures (Barrett et al., 2013; Maniscalco & Lau, 2012) are the current gold-standard in measures of metacognition. M*eta-d'* is the type 1 sensitivity that would be expected from the SDT-optimal observer, given the type 2 level information. In other words, *meta-d’* answers the question ‘what level of type 1 sensitivity would the optimal SDT observer need in order to obtain this confidence-accuracy relationship?’ Because *meta-d’* is in *d’* units, it can be compared to empirically observed values of *d’* to quantify how suboptimal the participant is, or how this sub-optimality is changed by experimental manipulations. The difference between *meta-d’* and *d’* has a clear interpretation in units that correspond to the standard deviation of the noise distribution. Type 2 *D’* on the other hand is formulated in different units from type 1 *d’*, making it hard to compare these two measures directly.

One assumes the participant has optimal metacognitive performance if *meta-d'* is equal to *d'*. Like the optimal SDT observer, they are using all of the evidence available. If *meta-d’* is lower than *d’*, the optimal observer could show the same degree of metacognitive accuracy shown empirically, but will use less type 1 information than the participant had. The ideal observer therefore outperforms the subject and the subject’s metacognitive accuracy is suboptimal. It is assumed that *meta-d'* will never be higher than *d',* as this would suggest the participant performed "super-optimally". In practice, this would support a model in which the observer has more information when making the type 2 decision than when making the type 1 decision, for example, after having had feedback on the type 1 decision, or having had to make a speeded type 1 decision.

There are several possible operational definitions of *meta-d’*, all of which rely on solving two pairs of equations, one pair obtained by considering type 2 performance following a positive type 1 response (e.g. “yes”) and the other obtained by considering type 2 performance following a negative type 1 response (e.g. “no”). All existing approaches fix the type 1 response bias (the relative type 1 threshold *c’*) to the empirical value, for the purposes of solving the equations for *meta-d’*. In general, the two pairs of equations cannot be solved simultaneously. Maniscalco & Lau (2012) adopt a data-driven approach, by proposing two methods for finding the best fit: minimising the sum of the squares of the errors leads to *meta-d'*$_{SSE}$, while maximum likelihood estimation leads to *meta-d'*$_{MLE}$.

Maniscalco and Lau's *meta-d'* formulation assumes symmetrical confidence thresholds. By contrast, meta-d' balance (*meta-d'*$_{b}$*,* Barrett et al., 2013) permits *response-conditional meta-d'* for positive and negative responses to differ. They propose this as a theory-driven rather than data-driven approach which affords an alternative calculation of *meta-d'*. They derive formulae for both positive and negative response-conditional *meta-d'*, but rather than solving these simultaneously, they take their mean solution, weighted according to the number of positive versus negative type 1 responses. Barrett et al. (2013) noted that the response-conditional *meta-d'* measures do not on their own provide stable, bias-invariant measures of metacognition; stability only comes when they are combined into a single measure.

Barrett et al. (2013) assessed how both *meta-d'*$_{b}$ and Maniscalco & Lau's *meta-d'*$_{SSE}$ behave under non-traditional SDT models. In practice, empirical data are messy and the paradigm may induce certain changes in how we envisage the statistical distributions of signal and noise. Importantly, Barrett and colleagues found that under an unequal variance model, even when departing from standard SDT (i.e. when the signal is enhanced or degraded between the type 1 and 2 levels, or when type 1 criterion is jittered across trials, representing fluctuations in attention) both versions remain relatively robust, especially when the type 1 threshold is varied. In these cases, however, *meta-d'*$_{b}$ seems slightly more consistent than *meta-d'*$_{SSE}$, which is unsurprising given that *meta-d'*$_{b}$ permits differences between the response-conditional metacognitive measures. When variances are equal, both measures are largely invariant to changes in type 1 and 2 thresholds under signal-degradation, signal-enhancement and criterion jitter.

Barrett et al. (2013) also looked at the behaviour of both *meta-d'* measures on finite data sets, and found that with small numbers of trials (approximately 50 trials per subject) both showed statistical bias and had higher variance than *d'*. However when 300 trials per subject were included in the analysis, bias approached zero and variance dropped substantially. Therefore to get the most out of these measures high numbers of trials per condition should be obtained.

The calculation of *meta-d'* is optimal when no type 1 or 2 hit or false alarm rate is too extreme, and not possible when any of these take the value zero or one. This leaves one with two possible sets of data exclusion criteria to consider. The `narrow exclusion criteria' only exclude a subject if any of the type 1 or response-conditional type 2 hit rates or false alarm rates are zero or one. These obviously maximise the number of subjects retained. An alternative choice is to use `wide exclusion criteria' which exclude subjects if any of the type 1 or response-conditional type 2 hit or false alarm rates lie at the extremities (<.05 or >.95). Simulations found narrow exclusion criteria to lead to greater variance of *meta-d'* but smaller bias than wide exclusion criteria.

In summary, both versions of *meta-d'* invert the calculation of type 2 performance from type 1 performance into a calculation of estimated type 1 performance given type 2 performance. Therefore, this method avoids many conceptual and theoretical problems related to computing an overall measure of metacognition. Moreover, these problems also seem to be avoided in practice. Although there is, as yet, no single, optimal computation for *meta-d'* it looks like $meta\text{-}d'_b$ is more robust to non-traditional SDT models whereas $meta\text{-}d'_{SSE}$ is less biased in small samples.

The main drawbacks of the *meta-d'* measures are that they are noisier than the alternative measures discussed above, and that response-conditional versions may be unstable. Nevertheless, these measures are most promising for capturing metacognition independently of response biases. In summary, these measures will give stable and meaningful results when sufficient trials are obtained and the standard assumptions of SDT hold to reasonable approximation.

3.8 Measuring consciousness using type 2 signal detection theory

The literature now offers robust measures of metacognition. So, how can we use measures of metacognition to deepen our understanding of consciousness?

There are arguments in the literature for using metacognition as a robust measure of visual awareness (Kunimoto & Miller 2001; Persuade et al 2007). These arguments claim that decision confidence taps in to the subjective states that underlie awareness. In many cases, it would indeed seem reasonable to assume that confidence will correspond with accuracy only when a target has been consciously perceived. However, this presumption was violated in blindsight patient GY. GY demonstrated above chance metacognition (Evans & Azzopardi, 2007), yet is clearly unaware of visual stimuli in the blind field (Persaud et al. 2007). Metacognition and awareness can also dissociate, such that metacognitive accuracy as measured by *meta-d’* is above chance for subliminally presented stimuli (Jachs, Blanco, Grantham-Hill, & Soto, 2015). Together, these results suggest that under certain circumstances we might (carefully) be able to use metacognition as a proxy measure of visual awareness or conscious knowledge. However, for a more rigorous assessment of unawareness we would hope to see a convergence with other measures that indicate unawareness – absence of EEG correlates such as the P300, for example.

There is a debate to be had about how we should interpret a metacognitive measure with relation to awareness. Imagine participants A and B take part in a psychophysical detection task. If A's *meta-d'* is twice that of B, are they "twice as aware" of the stimulus? Are they twice as *often* aware or twice as *likely to be* aware of the stimulus? When metacognition is at chance it is much easier to interpret the results in relation to awareness than when making relative judgements between above-chance values. By contrast, subjective confidence may tell us more about the subject’s experience of the stimulus. High confidence in a perceptual task corresponds to a high subjective probability of having correctly reported the target, whereas low confidence corresponds to uncertainty about the subject’s (task-relevant) perceptual content. Thus, these reports inform us about how the subject experiences the stimulus, rather than comparing this experience against the objective state of the world.

4

Prior Expectations Facilitate Metacognition for Perceptual Decision

The influential framework of 'predictive processing' suggests that prior probabilistic expectations influence, or even constitute, perceptual contents. This notion is evidenced by the facilitation of low-level perceptual processing by expectations. However, whether expectations can facilitate high-level components of perception remains unclear. We addressed this question by considering the influence of expectations on perceptual metacognition. To isolate the effects of expectation from those of attention we used a novel factorial design: expectation was manipulated by changing the probability that a Gabor target would be presented; attention was manipulated by instructing participants to perform or ignore a concurrent visual search task. We found that, independently of attention, metacognition improved when yes/no responses were congruent with expectations of target presence/absence. Results were modelled under a novel Bayesian signal detection theoretic framework that integrates bottom-up signal propagation with top-down influences, to provide a unified description of the mechanisms underlying perceptual decision and metacognition.

4.1 Introduction

Metacognition, or ‘cognition about cognition’, reflects the knowledge we have of our own decision accuracy and comprises an important, high-level component of decision making in both perceptual and cognitive settings. In perceptual decision, metacognition is often operationalised as the trial-by-trial correspondence between (objective) decision accuracy and (subjective) confidence. A key question in perceptual metacognition is how, and indeed whether, metacognition is affected by top-down influences such as attention and expectation. In the case of attention, it has long been known that it can improve visual target detection (Posner, 1980). However, the relationship between attention, confidence, and metacognition remains unclear. While Kanai and colleagues found that perceptual metacognition persists when attention is diverted (Kanai et al., 2010), other studies suggest that the absence of attention can lead to overconfidence (Rahnev et al., 2011; Wilimzig & Fahle, 2008).

Inspired by the growing influence of ‘predictive processing’ or ‘Bayesian brain’ approaches to perception and cognition (Clark, 2013; Gilbert & Li, 2013; Gilbert & Sigman, 2007; Summerfield & de Lange, 2014; Summerfield & Egner, 2009), empirical work on top-down attention is now complemented by a growing focus on the role of top-down expectations in decision making. In Bayesian terms, expectations can be conceived as prior beliefs that constrain the interpretation of sensory evidence. It has been shown that prior knowledge, either of stimulus timing ('when') or of stimulus features ('what'), facilitates low-level processing, as reflected in measures such as reaction time (Stefanics et al., 2010) and contrast sensitivity (Wyart et al., 2012). Such improvements are often accompanied by the attenuation of both the BOLD responses and ERP amplitude following expected relative to unexpected perceptual events (Egner et al., 2010; Melloni et al., 2011; Wacongne et al., 2011). As well as facilitating low-level perception, expectations may influence conscious content. This idea is supported by evidence for expectations inducing subjective directionality in ambiguous motion (Sterzer, Frith, & Petrovic, 2008) and lowering the threshold of subjective visibility for previously seen versus novel visual stimuli (Melloni et al., 2011). These effects are similar to those exerted by top-down attention. However, while it has been argued that attention and expectation reflect similar processes (Desimone & Duncan, 1995; Duncan, 2006), orthogonal manipulations of attention and expectation have demonstrated that, although they are tightly intertwined, they can have separable effects on neural activity (Hsu, Hämäläinen, & Waszak, 2014a; Jiang et al., 2013; Kok et al., 2011; Wyart et al., 2012).

One influential process theory within the Bayesian Brain framework is predictive coding (Beck, Ma, Kiani, & Hanks, 2008; Clark, 2013; Desimone & Duncan, 1995; Friston, 2009; Hohwy, 2013; Lee & Mumford, 2003). Predictive coding also posits that efficient processing is achieved by constraining perceptual inference according to the prior likelihood of that inference (‘expectations’). Here, the predictive models underlying perception are generally assumed to be multilevel and hierarchical in nature (Clark, 2013; Friston et al., 2012), incorporating priors related both to low-level stimulus features, and to high-level features representing object-level invariances. Plausibly, priors concerning subjective confidence for perceptual decisions may be implemented at high levels of the hierarchy. Based on this possibility, we set out to investigate whether the top-down influences of attention and prior expectation modulate perceptual metacognition.

To address whether expectation can improve metacognition we orthogonally manipulated both attention and expectation. This separated their effects, and was achieved by adopting a dual-task design. In a Gabor detection task, expectation was manipulated by informing participants of the probability of Gabor presence or absence as it changed over blocks. In this way, certain blocks induced an expectation of Gabor presence and others, of absence. In half of the blocks, participants were instructed to additionally perform a concurrent visual search task that diverted attention away from the detection task.

Objective performance can be assessed by using type 1 signal detection theory (SDT). By comparing signal type (e.g. present, absent) and response (present, absent), type 1 SDT enables a computation of independent measures of objective sensitivity and decision threshold (*d’* and *c*, respectively). We used type 2 SDT to assess metacognitive sensitivity (see Chapter 3). By obtaining trial-by-trial retrospective confidence ratings, metacognitive sensitivity and confidence thresholds can be computed from response accuracy and decision confidence. We used two such methods – type 2 *D’*, which is a direct analogue of type 1 *d’* (Kunimoto et al., 2001), and meta-*d’* (see Section 4.3.5.2 or Barrett, Dienes, & Seth, 2013; Maniscalco & Lau, 2012; Rounis, Maniscalco, Rothwell, Passingham, & Lau, 2010). Given that prior expectations have been shown to facilitate low-level processing, we hypothesise that expectations would also improve metacognitive sensitivity. We tested this hypothesis by considering the congruency between participants’ yes/no decision and the block-wise expectation of Gabor presence or absence. Specifically, we hypothesised that metacognitive sensitivity would be greater following expectation-congruent type 1 decisions (e.g. reporting target presence when expecting target presence), than following expectation-incongruent decisions (e.g. reporting presence when expecting absence).

4.2 Methods

4.2.1 Participants

Twenty-one participants (14 female) completed the experiment. All were healthy students from the University of Sussex, aged 18 to 31 (*M* = 20.4, *SD* = 3.2) and had normal or corrected-to-normal vision. The sample size for adequate power was computed using GPower 3.1 (Faul, Erdfelder, Lang, & Buchner, 2007), with estimated effect sizes derived from pilot studies. Data from one participant were excluded because their visual search task performance deviated by more than 1.5 *SD* from the mean (98.6% correct) and another, for having no variability in their confidence reports (100% confident). This left data from 19 participants for analysis, all of whom demonstrated, averaging over conditions, a Gabor detection *d’* and visual search accuracy that was within 1.5 *SD* from the mean. Participants were offered course credits for participating and informed, written consent was obtained. The experiment was approved by the University of Sussex ethics committee (C-REC).

4.2.2 STIMULI AND SETUP

Stimuli were generated using the Psychophysics toolbox for Matlab (Brainard, 1997; Kleiner, D., & Pelli, 2007) and presented on a 20 inch Dell Trinitron CRT display (resolution 1048x768; refresh rate 85 Hz). Participants were tested individually in darkened rooms and were seated 60cm away from the screen. Both stimuli and background were linearised using a Minolta LS-100 photometer (γ = 2.23607, Weibull fit). The background was greyscale and uniform.

4.2.3 DESIGN AND PROCEDURE.

This experiment implemented a novel dual-task design, which is depicted in figure 4.1. The critical task was to report the presence or absence of a near-threshold Gabor patch (which indeed, was either present or absent). The second task was a visual search task, in which it had to be determined whether a target (the letter 'T'), had been present or absent amongst distracters (letter 'L's).

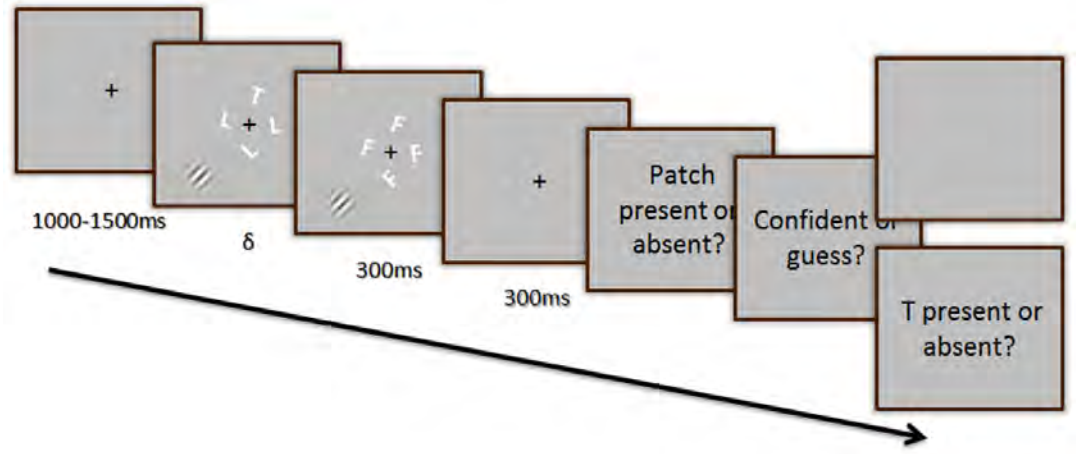

Figure 4.1. Trial sequence.

Across both staircases and experimental trials. In this trial, both the visual search and detection targets (T and Gabor, respectively) are present. Participants are prompted to respond to the visual search display in diverted attention trials (final, bottom) but not full attention trials (final, top). δ signifies the time that the visual search Ls and Ts were presented for. This time was titrated for each participant individually.

Trials began with the presentation of a white central fixation cross (0.38°x 0.38°, random duration between 500 and 1,500 ms). This was followed, on Gabor present trials only, by the appearance of the peripheral Gabor patch (spatial frequency 2c/°, Gaussian *SD* = 2°) in the lower-right quadrant of the screen. On each presentation, the phase was either 45° or 225° (50% chance of each). To reduce sensory adaptation effects, the precise location in which it was presented was jittered in both the horizontal and vertical direction from a baseline position of 25.2° x 21.08°. On each trial the jitter for each direction was randomly sampled from the interval [0.66°, 1.24°]. The contrast of the Gabor was titrated for each participant so that hit rate was 79.4% (see section 4.3.4, Staircases). In total it was presented for 388ms and had a gradual onset and offset. Immediately following the offset of the fixation cross, the central visual search array also appeared. On Gabor present trials, the Gabor and the array were therefore presented simultaneously. The array consisted of four white letters (1.43° x 1.43°) – either 3 ‘L’s and a ‘T’ (visual search target present, 50% chance) or 4 ‘L’s (visual search target absent, 50% chance) - arranged around fixation at 0°, 90°, 180° and 270°.

Trial-by-trial, the orientation of each letter took a random value between 0° and 359°. The time for which the letters remained on-screen was adjusted for each participant so that visual search percent correct was 79.4% (*M* = 254 ms, *SD* = 75 ms. See section 4.3.4, Staircases). To ensure that the task was difficult enough to divert attention, the array of letters was backwards-masked by an array of ‘F’s that remained on screen for 300ms. This masking array was followed by a series of on-screen response prompts, requesting un-speeded, key-press responses to: first, the Gabor task (Gabor present or absent); second, binary confidence in the accuracy of that report (confident or guess); finally, and in diverted attention conditions only (see next paragraph), the visual search task (T present or absent).

Expectation was manipulated in the Gabor task by changing the probability that it would be present versus absent over blocks of trials (25%, 50% or 75% probability of target presence). In the 25% condition, where Gabor presence was unlikely, an expectation of absence was induced. The 50% condition was a control, and in the 75% condition an expectation of presence was induced. Orthogonally to this expectation manipulation, attention was manipulated over blocks of trials by instructing participants to either perform or ignore the concurrent visual search task. When participants were in a ‘perform visual search’ block, their attention was diverted from the critical Gabor detection task, whereas when they were instructed to ignore the visual search array, their attention was fully focused on Gabor detection. In the diverted attention condition, participants were instructed to prioritise the visual search task. Thus, each block was associated with an expectation of Gabor presence or absence and a degree of attentional resource for the Gabor task (full/diverted). Before each block began, both the probability of Gabor presentation and instructions to either perform both tasks or ignore the Gabor were presented on-screen. At the end of each block, if visual search accuracy had dropped below 60% on-screen feedback reminded participants to maintain their concentration on the visual search task. Participants completed 36 blocks in total (6 of each of the 6 conditions, counterbalanced) and each block had 12 trials. This gave a total of 432 trials.

Before data collection began, instructions for the tasks were presented on-screen. The on-screen instructions were additionally read to the participant to ensure that they were fully understood. These explained that the probability of target presentation in the upcoming block would be given (25%, 50% or 75%) and that the information was correct and would help them complete the difficult task. Participants were instructed to fixate centrally throughout and to be as accurate as possible in all of their (un-speeded) responses. Next, participants completed a set of practice trials for each type of task (staircases and experimental conditions). Next, three psychophysical staircase procedures were completed (see section 4.2.4) and finally, the experimental trials. Once all experimental trials had been completed, participants were debriefed.

4.2.4 STAIRCASES.

We required performance in the Gabor detection task to be equated across levels of attention and across participants. Furthermore, the difficulty of the visual search task also had to be controlled across participants. To this end, three adaptive staircase procedures were completed prior to the experimental trials. The first staircase adjusted Gabor contrast under full attention, the second, the time for which visual search ‘L’s and ‘T’s were presented and the third, Gabor contrast under diverted attention. The staircases set performance (percentage correct for the visual search task and hit rate for the Gabor task) in each task at 79.4%. Each of the three procedures consisted of two interleaved, identical staircases, which terminated after 8 reversals. The visual display was identical to that in experimental trials (see section 4.2.3 and figure 4.1), however the reports requested from participants varied across procedures. During these procedures, confidence judgments were not requested and there was a 50% chance of Gabor presentation.

In staircase 1, Gabor detection was performed under full attention (i.e. ignore visual search). Participants were instructed to fixate centrally, ignore the visual search display and report peripheral Gabor presence or absence. The initial contrast of the attended target Gabor was 5% and this was titrated by the staircases. The (ignored) visual search ‘L’s and ‘T’s were presented for 300 ms before they were masked.

In staircase 2, the visual search task was performed but the Gabor task was not. Participants were instructed to ignore the Gabor and only perform the visual search task. Here, they reported whether a target ‘T’ was present or absent in an array of distracter ‘L’s. The visual search array of ‘L’s and ‘T’s were initially presented for 300 ms before being masked, and this duration was titrated by the staircases. The (ignored) Gabor, if present, had the contrast determined in staircase 1.

In staircase 3, both tasks were performed. Participants were instructed to prioritise the visual search task while concurrently performing the Gabor detection task. Visual search letters were presented for the duration determined in staircase 2. The Gabor was initially presented at 1.05 times the contrast level acquired in staircase 1. The contrast of the unattended Gabor was titrated over the course of the procedure. Participants responded to the Gabor task first and the visual search task second (as in the experimental trials). If participants' mean visual search accuracy across the staircase dropped below 60% they received on-screen instructions to maintain concentration on the visual search task.

4.2.5 Analysis

4.2.5.1 Statistical analyses.

Objective detection performance for the Gabor detection task was assessed using type 1 signal detection theory (SDT; Green & Swets, 1966) measures *d'* (detection sensitivity) and *c* (decision threshold). A negative/positive *c* reflects a bias towards reporting target presence/absence. Visual search performance was also assessed using *d'* and *c*. Because we required *d'* and *c* values to remain independent of each other, adjusted type 1 *d'* was not used. Unless otherwise stated, alpha is set at 5%, the assumption of sphericity has been met and post-hoc tests are FDR corrected (Banjamini & Hochberg, 1995) throughout.

4.2.5.2 Type 2 Signal Detection Theory.

Metacognitive sensitivity was measured by obtaining trial-by-trial confidence ratings and using type 2 SDT to assess the relationship between confidence and accuracy (Barrett, Dienes, & Seth, 2013; Galvin, Podd, Drga, & Whitmore, 2003; Kunimoto et al., 2001; Macmillan & Creelman, 2004). Type 2 measures are calculated analogously to the type 1 case: type 2 hits (correct and confident) and correct rejections (incorrect and guess) are compared with type 2 misses (correct and guess) and false alarms (incorrect and confident). From these, type 2 *D'* (metacognitive sensitivity) and type 2 *C* (confidence threshold) can be computed (Kunimoto et al., 2001). Type 2 hit rate (HR) and type 2 false alarm rate (FAR) are calculated as follows (where the subscript ‘2’ indicates type 2 SDT outcomes):

$$HR = \frac{\sum H_2}{\sum H_2 + \sum M_2}\,, \qquad FAR = \frac{\sum FA_2}{\sum FA_2 + \sum CR_2}$$

Thus, HR reflects confidence for correct responses and FAR reflects confidence for incorrect responses.

Type 2 *D’* and type 2 *C* are defined as:

$$D' = Z(HR) - Z(FAR), \qquad C = \frac{-(Z(HR) + Z(FAR))}{2}$$

where Z is the standard Z-score, i.e. the inverse cumulative density function of the standard normal distribution. To distinguish type 2 variables from their type 1 counterparts we denote type 1 variables in lower-case (e.g. type 1 *d’*) and type 2 in upper case (e.g. type 2 *D’*).

It is known that type 2 *D’* is highly biased by both type 1 and 2 thresholds (Barrett et al., 2013; Evans & Azzopardi, 2007; Galvin et al., 2003). An alternative measure is the ‘bias-free’ meta-*d’.* This is an estimate of the type 1 *d’* an SDT-optimal observer would need to have to generate the type 2 performance shown (for an in-depth explanation see Barrett et al., 2013 or Maniscalco & Lau, 2012). Importantly, meta-*d’* is measured in the same units as *d’*. This permits a direct comparison between objective and subjective sensitivity. Considering meta-*d’* as a proportion of *d*’ gives us metacognitive efficiency, or the amount of type 1 information that is carried forward to the type 2 level. To take advantage of this feature we additionally analysed our results using meta-*d’/d'.* We calculated meta-*d’*-balance from freely available online code (Barrett et al., 2013). This calculation was supplemented by a maximum likelihood estimation of SD_{noise}:$SD_{signal+noise}$ from the group-level data, also using freely available online code (columbia.edu/~bsm2015/type2sdt; Maniscalco & Lau, 2012).

As described in the introduction, we hypothesised that metacognitive performance would be improved when type 1 decisions are based on prior expectations. Testing this hypothesis requires comparing decisions that were based on (i.e. congruent with) prior expectations with those that were not. In the 25% condition, target absence is most probable meaning that an ‘absent’ report would be expectation-congruent and a ‘present’ report would be incongruent. The opposite would be true for the 75% condition. We therefore computed, for each condition, type 2 *D’* following ‘present’ responses (hits and false alarms) and type 2 *D’* following ‘absent’ responses (misses and correct rejections). Analogous response-conditional meta-*d’* estimates were obtained from freely available online code (see Barrett et al., 2013, supplementary materials). Unfortunately, response-conditional meta-*d'* is unlikely to be robust to criterion shifts like its response-unconditional counterpart (Barrett et al., 2013).

For all type 2 measures, a significant response by expectation interaction would demonstrate an effect of congruency. Note that we could not use a standard (i.e. response-unconditional) *D'* or meta-*d'* measure, because in this case degraded metacognition following one response could cancel out the improved metacognition following the alternative response.

4.3 Results

4.3.1 Expectation can be separated from attention

To verify that the concurrent visual search task successfully manipulated attention we compared the contrast thresholds obtained in the full and diverted attention staircases. As expected, a one-tailed paired *t*-test revealed a significant increase in contrast in the dual-task (*M* = 0.080, *SE* = 0.011) relative to the single-task (*M* = 0.032, *SE* = 0.002) conditions, bootstrapped *t*(18) = 4.64, *p* = .001, 95% CI = [-0.06, -0.03], dz = 1.06. Thus, the paradigm successfully manipulated attention.

Next, the effects of expectation and attention on each of (Gabor) detection sensitivity *d'* and (Gabor) decision threshold *c* were examined. These analyses addressed three questions: first, whether *d'* had been successfully equated across levels of attention and expectation; second, whether the expectation manipulation successfully biased *c*; third, whether expectation and attention were successfully separated at the type 1 level (i.e. did not interact under *d'* or *c*).

First, we performed a repeated-measures Expectation (0.25, 0.5, 0.75) x Attention (full, diverted) analysis of variance (ANOVA) on type 1 *d'*. This revealed that sensitivity did not significantly differ across the full (*M* = 2.39, *SE* = 0.16) and diverted (*M* = 2.00 *SE* = 0.18) attention conditions, *F(*1, 18) = 3.03, *p* = .099, η_p^2 = .144, or across Expectation conditions ($M_{.25}$ = 2.15, $SE_{0.25}$ = 0.11, $M_{.50}$ = 2.28, $SE_{.50}$ = 0.15, $M_{.75}$ = 2.14, $M_{.75}$ 0.13), *F(*2, 36) = 2.12, *p* = .124, $\eta_p{}^2{}_E$ = .101 (Figure 4.2 A). Type 1 sensitivity was therefore successfully equated across all six conditions. This means that any changes in type 2 sensitivity cannot be attributed to changes in the amount of type 1 information. There was no significant interaction between Attention and Expectation under *d', F*(2, 36) = 1.12, *p* = .34, η_p^2 = .059, suggesting that the two factors were successfully separated with respect to type 1 detection performance.

A repeated-measures Expectation (0.25, 0.5, 0.75) x Attention (full, diverted) analysis of variance (ANOVA) under decision threshold *c* revealed a significant main effect of Expectation, *F*(2, 36) = 9.18, *p* = .001, η_p^2 = .338. A trend analysis demonstrated that decision threshold linearly liberalised (more likely to report target present) as the probability of target presence increased, *F*(1, 18) = 15.72, *p* =.001, η_p^2 = .466. The paradigm therefore successfully manipulated expectation. Attention had no significant main effect on decision threshold, *F*(1,18) = 3.14, *p* =.093, η_p^2 = .148 and did not significantly interact with Expectation, *F*(2,36) = 0.85, *p* = .434, η_p^2 = .045 (Figure 4.2B). Therefore, attention and expectation were separated with respect to type 1 decision threshold, as well as type 1 sensitivity.

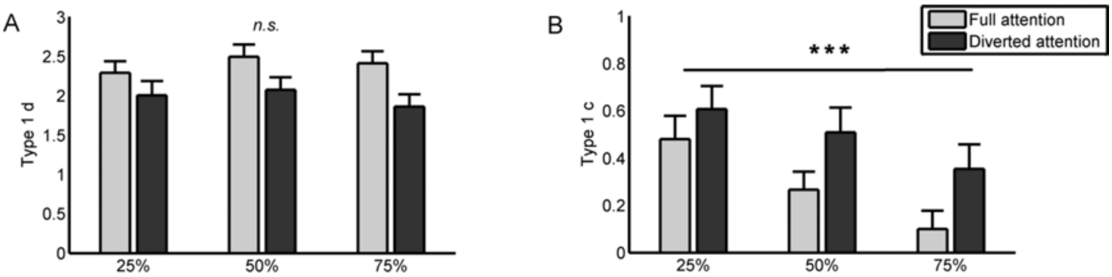

Figure 4.2. Effects of expectation and attention on sensitivity and bias

A. Type 1 *d'* as a function of expectation and attention. **B.** Type 1 criterion c as a function of expectation and attention. Error bars represent within-subjects SEM. *** *p* < .001, ** *p* < .01, * *p* < .05, *n.s*. non-significant.

In the diverted attention condition, participants were instructed to perform the detection and the visual search task simultaneously, prioritising visual search. However, if participants were unable to divide their attention across the two tasks then we would expect a significant negative correlation between trial-by-trial Gabor detection and visual search accuracy. To address this question we computed the Spearman's correlation coefficient between trial-by-trial detection accuracy scores on the two tasks for each participant. A one-sample bootstrapped *t*-test against zero revealed that at the group-level there was no significant trade-off in performance between the two tasks, *M* = 0.02, *SD* = 0.09, *t*(18) = 0.94, *p* = .361, 95% CI [-.023, .059]. Thus, participants were able to perform the two perceptual tasks simultaneously.

Participants were able to perform the tasks simultaneously, but if the visual search task interfered with Gabor detection sensitivity, we might expect a significant negative correlation between experiment-wise performance in the two tasks. To address this concern we calculated *d'* and *c* for the visual search responses and correlated them with their Gabor detection counterparts. Across participants there was no significant (Pearson's) correlation between visual search *d'* and (diverted attention) Gabor *d', r*(19) = .250, *p* = .302, bootstrapped 95% CI [-.326, .623]. Similarly, there was no significant (Pearson's) correlation between type 1 decision thresholds for the two tasks, *r*(19) = .359, *p* = .131, bootstrapped 95% CI [-.043, .723]. These results suggest that performing the visual search task did not significantly interfere with performing the Gabor detection task. This, combined with the absence of a negative correlation between trial-by-trial accuracy on the two tasks and with the absence of attention by expectation interactions under *d'* and *c*, demonstrates that attention and expectation were sufficiently separated at the type 1 level.

The results so far indicate that the paradigm successfully influenced both expectation (participants were more likely to report target absence when the probability of target presentation was low than when it was high) and attention (contrast sensitivity was reduced when attention was diverted). Furthermore, they indicate that expectation and attention did not significantly interact. Given this, we were able to examine how metacognitive sensitivity is specifically affected by expectation and attention, without confounds of task difficulty.

4.3.2 EXPECTATION IMPROVES METACOGNITIVE PERFORMANCE

Our main hypothesis was that metacognition would be improved following an expectation-congruent response. In the 25% condition, where target absence is expected, misses and correct rejections ('no') would be expectation-congruent responses and false alarms and hits ('yes') would be incongruent. The reverse is true for the 75% condition, where target presence is expected.

To test our hypothesis, response-conditional type 2 *D’s* (see Methods) were subjected to a repeated-measures Expectation (0.25, 0.5, 0.75) x Attention (full, diverted) x Report (present, absent) analysis of variance (ANOVA).

Critically, the ANOVA revealed a significant two-way interaction between Expectation and Report, $F(2,36) = 5.60$, $p = .008$, $\eta_p^2 = .238$. To further probe this effect we collapsed across attention conditions and performed *a priori* trend analyses. *D’* for target present reports exhibited a significant linear trend with Expectation, $F(1,18) = 13.85$, $p = .001$ (1-tailed), $\eta^2 = .435$ such that as the probability of target presentation increased from 25% (target presence improbable) to 75% (target presence probable), type 2 *D’* increased (Figure 4.3A). Similarly, when participants reported the Gabor as absent there was a significant linear trend with Expectation in the opposite direction, $F(1,18) = 3.83$, $p = .033$ (1-tailed), $\eta^2 = .175$: as the probability of target presentation

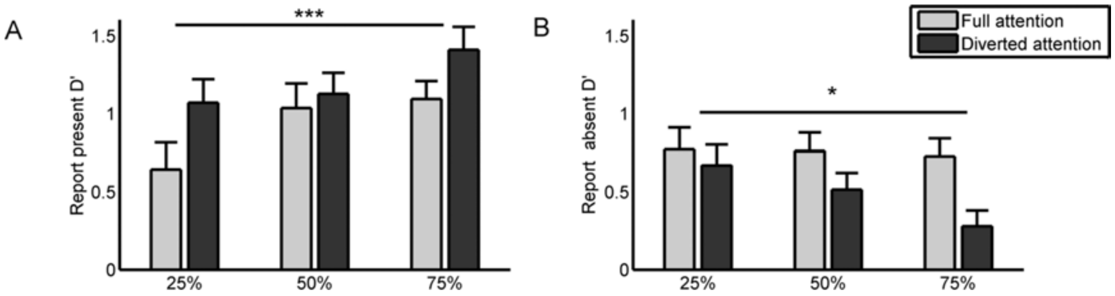

Figure 4.3. Response-conditional type 2 D' as function of expectation and attention.

Black lines indicate linear changes in D' with expectation, independently of attention. **A.** Type 2 D' for reports of target presence increases with expectation of presence **B.** Type 2 D' for reports of target absence increases with expectation of absence. Error bars are with-subjects SEM. * *p* < .05 ** *p* < .01, *** *p* < .001.

decreased from 75% (target absence improbable) to 25% (target absence probable), type 2 D' increased (Figure 4.3B). This congruency effect supports our hypothesis that expectation improves metacognition.

As well as a significant Report x Expectation interaction, there was a significant interaction between Report and Attention, *F*(1,18) = 5.61, *p* = .029, η_p^2 = .238. This interaction was driven by the presence of a significant difference between *D'* for absent and present reports under diverted attention (*M* = 0.49, *SE* = 0.13 and *M* = 1.20, *SE* = 0.19, respectively) , *F*(1,18) = 6.32, *p* = .022, η^2= .260, but not under full attention (*M* = 0.75, *SE* = 0.11 and *M* = 0.92, *SE* = 0.16, respectively), *F*(1,18) = 0.84, *p* = .372, η^2= .045. This unexpected result suggests that inattention impairs metacognition for unseen but not seen targets. The ANOVA did not reveal a significant main effect of Expectation on *D',* *F*(2,36) = 0.64, *p* = .533, η_p^2 = .034. This is unsurprising, because the influence of expectation is seen by comparing expectation-congruence relative to incongruence. There was also no significant main effect of Attention on type 2 *D', F*(1,18) = 0.01, *p* = .953, η_p^2 = .001, and no significant Report by Attention by Expectation interaction, *F*(1.60,28.81) = 0.11, *p* = .858, η_p^2 = .006 (ε = .748, Huynh-Feldt corrected).

In summary, these data under type 2 *D'* indicate that metacognitive performance improved when reports of target absence or presence were congruent with participants' expectation (25% or 75% condition, respectively), as compared to when they were incongruent (75% or 25% condition respectively).

4.3.3 Expectation liberalises confidence judgments

Given that expectation improved metacognitive performance, did expectations also increase subjective confidence? Type 2 confidence threshold can be interpreted as a proxy measure of the strength of the perceptual experience (Fleming & Lau, 2014). We therefore asked whether expectation-congruent reports were associated with higher confidence ratings than their incongruent counterparts. Such a result could be interpreted as expectations strengthening the associated perceptual experience.

We tested this possibility by asking whether expectation and report interacted under confidence threshold *C.* Confidence threshold is analogous to type1 decision threshold, signalling over-confidence when it is negative and under-confidence when it is positive. Therefore, if expectation liberalises confidence judgments we would expect confidence thresholds for 'present' responses to liberalise with increased expectation of presence. Following an 'absent' response, we would expect confidence to liberalise with increasing expectation of target absence (i.e. *decreasing* expectation of target presence).

To test this possibility we ran a repeated-measures Expectation (0.25, 0.5, 0.75) x Attention (full, diverted) x Report (present, absent) analysis of variance (ANOVA) on *C*. This revealed a significant three-way interaction, $F(2,36) = 4.69$, $p = .015$, $\eta_p^2 = .207$, which was not found in the ANOVA on type 2 *D’*. We analysed this interaction by performing simple effects analyses separately for the full and diverted attention conditions. Under full attention, Report and Expectation significantly interacted, $F(2,36) = 15.95$, $p < .001$, $\eta_p^2 = .470$. The pattern was the same as that found under type 2 *D’*. With increasing probability of target presence, there was a linear *decrease* in type 2 *C* (more likely to report high confidence) when the target was reported as present, $F(1,18) = 11.48$, $p = .002$, (one-tailed) $\eta^2 = .272$. However there was a linear *increase* in type 2 *C* (more likely to report low confidence) when the target was reported as absent, *F*(1,18) = 25.29, $p < .001$ (one-tailed), η^2= .584. Thus, expectations liberalise subjective confidence judgments under full attention.

By contrast, under diverted attention there was neither a significant main effect of Expectation, *F*(1,18) = .339, η_p^2 = .051, nor a significant interaction between Expectation and Report, *F*(2,36) = 2.84, *p* = .082, η_p^2 = .136.

The ANOVA under *C*, revealed no significant main effect of Attention, *F*(1,18) = 0.83, *p* = .374, η_p^2 = .044, and no significant interactions between Attention and Report, *F*(1,18) = 4.09, *p* = .058, η_p^2 = .185, or Attention and Expectation *F*(1,18) = 0.83, *p* = .444, η_p^2 = .044.

While type 2 *C* quantifies confidence relative to accuracy, the probability of correct and incorrect responses cannot be represented as Gaussian distributions over any transformation of the type 1 axis (Galvin et al., 2003). This means that we cannot determine whether expectations change the *evidence* needed to report decisions with high confidence, or whether confidence changes because of bias at the level of objective report. To address this point, confidence thresholds for ‘yes’ and ‘no’ reports over the type 1 axis were computed (see Chapter 3). These thresholds are placed over the type 1 decision axis and reflect the evidence required to report decisions with high confidence. We divided each threshold by its respective *d’* to account for differences in sensitivity when making confident versus guess decisions. We denote these normalised confidence thresholds for yes and no responses C_1’+ and C_1’- respectively. To determine the separation of each threshold from the type 1 decision threshold we took the log absolute distance between each of C_1’+ and C_1’- and *c’ = c/d’*. These values quantify the separation, in type 1 evidence units, between the confidence threshold and decision threshold, and in turn reflect metacognitive bias.

Metacognitive bias was subjected to an Expectation (0.25, 0.5, 0.75) x Attention (full, diverted) x Report (present, absent) analysis of variance which revealed a significant Expectation by Report interaction, *F*(2,36) = 4.93, *p* = .013, η_p^2 = .215. As shown in figure 4.4, metacognitive bias for ‘yes’ responses linearly

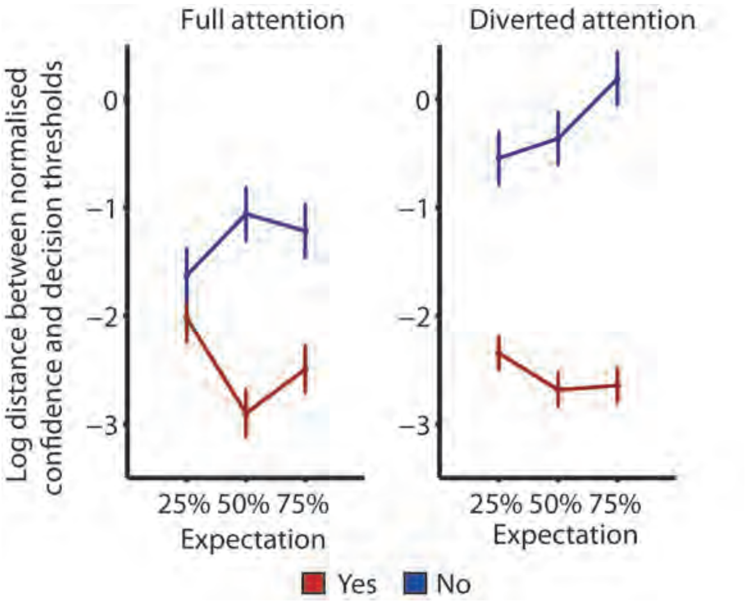

Figure 4.4. Response-specific confidence thresholds C_1'+ (red) and C_1'- (blue) as a function of attention and expectation.

Under full attention (left), the normalised log distance between the type 1 decision threshold and the confidence threshold for 'yes' responses (red) becomes smaller as target presence is increasingly probable. Simiarly, this distance becomes smaller for 'no' responses (blue) as target absence is increasingly probable. This effect of expectation is lessened under diverted attention.

liberalised with expectation of presence, $F(2,36) = 5.21$, $p = .010$, $\eta_p^2 = .224$, and for 'no' reports, marginally liberalised with increasing expectation of absence, $F(2,36) = 2.84$, $p = .082$, $\eta_p^2 = .136$. There was a marginal three-way interaction ($p = .079$) such that, consistent with results under type 2 *C*, congruence shaped confidence under full attention ($p = .010$) but only marginally so under diverted attention ($p = .063$). Together, these results indicate that expectations biased confidence over and above effects on type 1 decision, such that expected percepts may have required less evidence to be reported with high confidence.

Summarising so far, metacognition improved for expectation-congruent perceptual decisions, independently of whether attention was focused on or diverted from the task. This effect was mirrored under confidence thresholds, but primarily under full attention. Under these conditions, the perceptual experience associated with expectation-congruent decisions may be stronger than that for expectation-incongruent decisions.

4.3.4 REPORT-EXPECTATION CONGRUENCY INCREASES *META-D’*.

To assure the robustness of our findings under type 2 *D’*, we re-analysed the data using response-conditional meta-*d’*. As mentioned in section 2.5.2, given the type 2 performance observed, meta-*d’* is the type 1 *d’* that would be expected from the SDT-optimal observer who used all of the available type 1 information. Meta-*d’*/*d’* is therefore the proportion of type 1 information used in the type 2 decision. We expected to find the same pattern of results as those obtained under *D’* – a Report by Expectation interaction whereby meta-*d’*/*d*’ increases with response-expectation congruency. Only 1/19 of our participants fully met the criteria for assuring reliable meta-*d’* estimates (for all 6 conditions, 0.05 ≤ hr, far, HR_+, FAR_+, HR_-, FAR_- ≤ 0.95; see Barrett et al., 2013). We therefore retained participants who met these criteria in at least 3/6 conditions. This left us with 12 participants for the analysis.

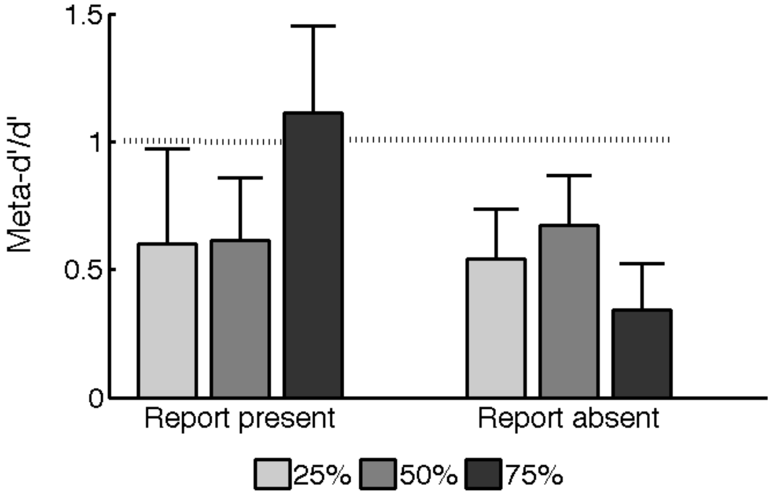

Figure 4.5. Meta-d’/d’ as function of expectation and type 1 report.

Significant interaction between expectation and report, where meta-*d’*/*d’* increases with expectation-response congruence. Error bars are with-subjects SEM. * *p* < .05, ** *p* < .01, *** *p* < .001.

As for the previous analyses, a repeated-measures Expectation (0.25, 0.5, 0.75) x Attention (full, diverted) x Report (present, absent) analysis of variance (ANOVA) was conducted, but this time using meta-*d'*/*d* as the dependent variable*.*

Consistent with our previous result, the analysis revealed a significant Expectation x Report interaction, *F*(2,22) = 8.75, *p* = .002, η_p^2 = .443. *A priori* trend analyses revealed that following a ‘present’ response, meta-*d'*/*d'* linearly increased with expectation of target presence, *F*(1,11) = 5.12, *p* = .022 (one-tailed), η^2 =.318. Following an ‘absent’ response there was a significant decrease in meta-*d'*/*d'* as the probability of target presence increased, *F*(1,11) = 4.22, *p* = .032 (one-tailed), η^2 =.277. These patterns are illustrated in figure 4.5. We found no other significant main (all *F* < 2.37, all *p* > .15, all η_p^2 < .29) or interaction (all *F* < 0.99, all *p* > .32, all η_p^2 < .09) effects. This pattern of results held under slightly narrower and broader exclusion criteria (i.e. proportion of stable conditions).

Summarising, report-expectation congruency improves metacognitive performance when measured by response-conditional meta-*d'*, as well as when measured by response-conditional *D'*.

4.3.5 A Type 2, Bayesian Signal Detection Theoretic Model of Expectation and Top-Down Attention

To model the influence of top-down expectation on metacognitive sensitivity we extended standard signal detection theory (SDT) to incorporate prior expectations (Figure 4.6). In our model, the evidence is the internal variable X in SDT (the internal representation of Gabor contrast) and the expectation is the probability of Gabor patch presentation. The ‘signal’ and ‘noise’ distributions were reformulated as posterior distributions of the cases of target present and absent, given both the evidence X *and* the expectation*.* Therefore the primary effect of expectations is to change the internal representation of the stimulus (the probability distributions), which in turn will induce apparent shifts in type 1 and 2 criteria.

Type 1 and 2 decision criteria (*c* and *C*) were formulated as distinct thresholds for the posterior ratio of probabilities of present (S=1) to absent (S=0). For probability *p* of stimulus present and evidence *x*, this ratio, which we denote by R, is given by

$$R = \frac{P(S=1|x)}{P(S=0|x)} = \frac{P(x|S=1)P(S=1)}{P(x|S=0)P(S=0)} = \frac{\varphi_{d',\sigma}(x)\times p}{\varphi_{0,1}(x)\times(1-p)}$$

where $\varphi_{\mu,s}$ is the probability density function of a normal distribution with mean μ and standard deviation *s*. Assuming the SDT model, this ratio monotonically increases with the evidence *x*.

To model the effect of diverted attention we implemented the solution proposed by Rahnev et al. (2011), in which inattention increases the trial-by-trial internal noise. To assess whether this model could account for our data we computed the response-conditional type 2 *D'*s predicted by the model at varying, continuous levels of prior expectation of patch present. This was done separately for the full and diverted attention cases.

Parameters were determined in the following way: Type 1 *d'* was set to 2.39 and 2.00 for the full and diverted attention conditions respectively, reflecting the mean empirical values we obtained.

For each level of attention, the type 1 and 2 thresholds for *R* were based on the mean empirical type 1 and 2 hit and false alarm rates in the respective 50% expectation condition. For the full attention case, the obtained type 1 threshold was *R* =1.88, and the upper and lower type 2 thresholds were *R* = 4.27 and *R* = 0.68 respectively.

For the diverted attention case, these were respectively *R* = 2.52, *R* = 4.06 and R=0.86. For full details on obtaining type 1 and 2 decision thresholds from type 1 and 2 hit and false alarm rates, see Barrett et al. (2013). Notice that, since contrast was increased in the experiment for diverted attention, the models for full and diverted attention were approximately the same; only the threshold values (*R*) differed slightly.

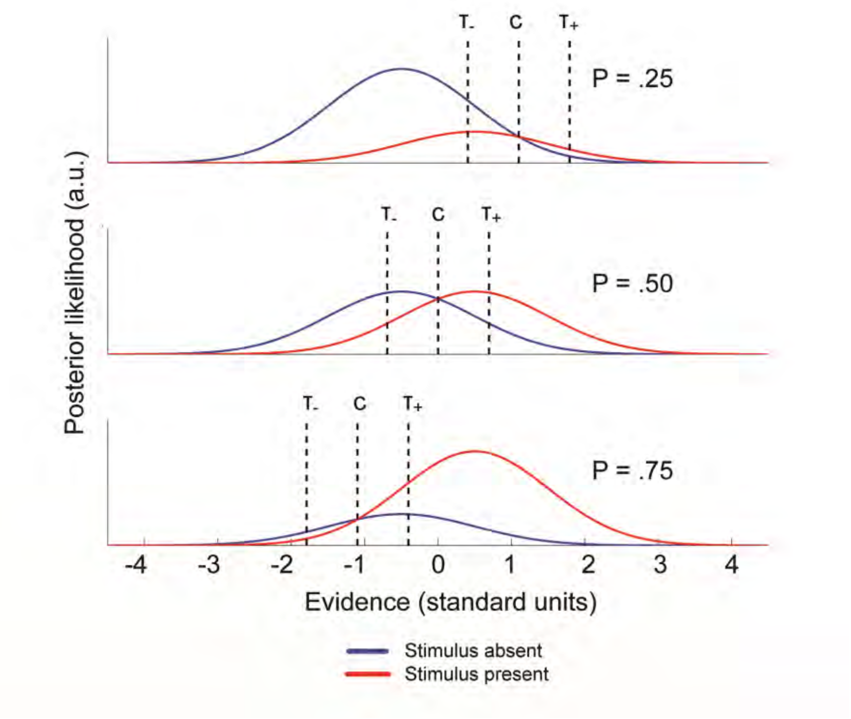

Figure 4.6. A Bayesian signal detection theoretic model of prior expectation.

Each panel plots the posterior likelihood of a perceptual event against the evidence given distinct prior probabilities (P) of stimulus present. The blue curve represents the event of stimulus absence and the red curve, stimulus presence. Type 1 d' (the distance between the blue and red Gaussians) is held at 1. The curves are aligned so that criterion is unbiased when *p* = .50. The dashed lines show the decision (c) and confidence (τ_+, τ_-) thresholds. These are each determined by a fixed posterior likelihood ratio R for stimulus present to stimulus absent. These plots illustrate that detection, as well as confidence about detection, liberalises with increased prior expectation on Bayesian SDT.

Figures 4.7A-D compare the predicted and empirical *D'*s across levels of report and attention. In agreement with the empirical data, predicted *D'* for positive responses increased with prior expectation for target present (Figures 4.6A and 4.6B), while *D*' for negative responses, it decreased (Figures 4.6C and 4.6D). As was the case for the empirical results, this decrease demonstrates an increase in *D*' with increased prior expectation for target absent. The model predicted slight attentional modulations of *D*', which reflect numerical differences in empirical type 1 *d*'.

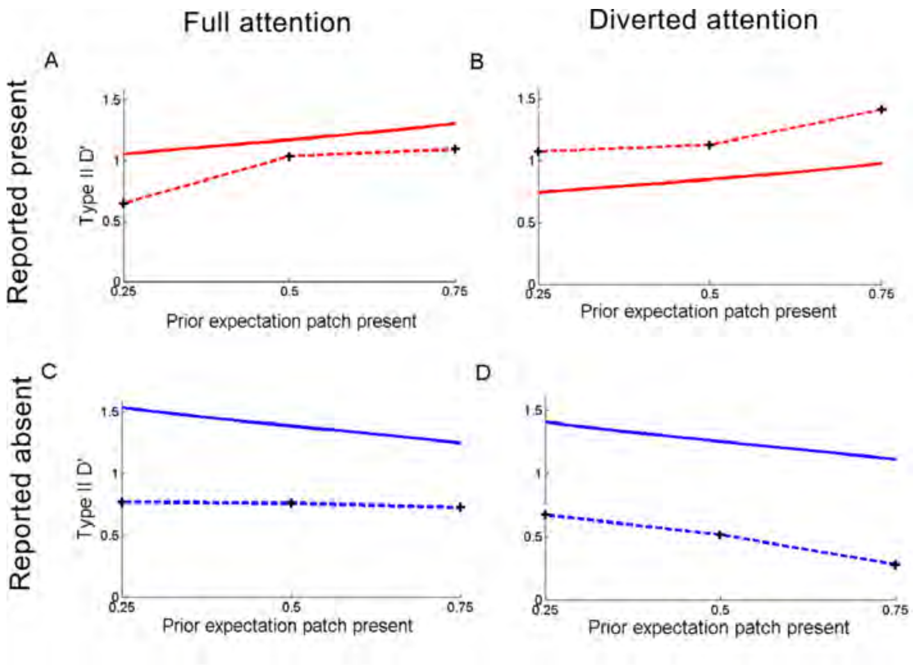

Figure 4.7. Modelling of empirical results.

Solid lines represent stimulated results over continuous probabilities of target present. Dashed lines are the corresponding empirical results collected over 25, 50 and 75% of target presence. The top and bottom rows show results for reported present and absent trials, respectively. The leftward and rightward columns show results for full and diverted attention

Simulated *D*' values for 'absent' responses also took substantially higher values than those collected empirically. Moreover, simulated *D'* was higher for absent than for present responses, whereas the reverse trend was found empirically.

These two features persisted for variant models on which signal and noise distribution variances were unequal. They are likely attributable to asymmetries in the degradation of type 1 evidence available for metacognition, an investigation of which is beyond the scope of this chapter. In summary, our modelling analyses demonstrate that the observed dependencies of metacognitive performance on prior expectation are consistent with a signal-detection theory model extended according to Bayesian principles to incorporate expectations as priors.

4.3.6 Effect of expectation on a concurrent visual search task

So far we have shown that expectations of Gabor presence or absence improve metacognition for the Gabor detection task. Given this, could expectations of

Gabor target presence or absence also facilitate perceptual decisions for the visual search task? The expectations induced by the paradigm pertained to the Gabor target, however the influence of these expectations may free perceptual and cognitive resources for other tasks.

To address this question, we first asked whether expectation affects decisions made on the visual search task (i.e. T presence or absence). This was achieved by computing type 1 *c* for the visual search task as a function of expectation. Visual search data from the full attention condition could not be analysed because the required responses were not collected.

A one-way Expectation (.25, .50, .75) repeated measures ANOVA under visual search criterion c_{vs} revealed a significant effect of Expectation, *F*(2,36) = 6.17, *p* = .005, η^2 = .255. However, rather than expectation of Gabor presence inducing a liberal criterion shift under the visual search task, as it did under the Gabor task, there was a significant quadratic trend, *F*(1,18) = 11.74, *p* = .003, η^2 = .395. This trend was such that participants were more likely to report that a T was present (liberal shift) in the 50% condition (*M* = 0.19, *SE* = 0.09) than when they had a task-irrelevant prior expectation of Gabor presence or absence (25% and 75% conditions, *M* = 0.35, *SE* = 0.09, *M* = 0.32, *SE* = 0.08, respectively). Therefore the task-irrelevant expectation of Gabor presence or absence did not bias participants towards reporting presence or absence on the visual search task. Rather, expectations induced a conservative shift in *c* relative to the neutral (50%) condition.

Given that expectation of Gabor presence or absence biased decisions made in the visual search task, they may also have affected sensitivity. To test this, we calculated visual search *d*' as a function of Gabor detection accuracy and expectation-Gabor response congruence. The factor Congruence was formed by grouping trials according to whether the response to the Gabor task (present or absent) was congruent or incongruent with the prior expectation (75%, where they expect presence, 50%, which is neutral, 25%, where they expect absence). This factor represents the influence of expectation on Gabor decision. If visual search performance is modulated by the effect of expectation on the Gabor task then there should be an effect of this factor.

A repeated-measures Gabor accuracy (correct, incorrect) x Gabor congruence (incongruent, neutral, congruent) ANOVA on visual search d'_{vs} revealed a significant main effect of Gabor accuracy, $F(1,18) = 4.80$, $p = .015$, $\eta_p^2 = .288$, whereby d'_{vs} was higher following a correct ($M = 1.72$, $SE = 0.16$) than an incorrect ($M = 1.31$, $SE = 0.18$) response on the Gabor detection task. Therefore high perceptual sensitivity for the Gabor was associated with high perceptual sensitivity for the visual search task as well. The ANOVA also revealed a marginally significant interaction between accuracy and congruence, $F(2,36) = 2.95$, $p = .065$, $\eta_p^2 = .141$. Post-hoc trend analyses revealed that d'_{vs} linearly increased with expectation-Gabor response congruence following a correct response on the Gabor task, $F(1,18) = 4.49$, $p = .048$, $\eta^2 = .200$ and linearly decreased with congruency following an incorrect Gabor response, $F(1,18) = 5.27$, $p = .034$, $\eta^2 = .226$. This result suggests that visual search sensitivity improved when the (Gabor) expectation had been valid (i.e. met in the stimulus-conditional sense). This follows from the observation that the expectation was only valid in trials where correct and congruent or incorrect and incongruent responses were made. To illustrate, in the 25% condition, correct responses were correct rejections (congruent, valid expectation) or hits (incongruent, invalid expectation). The former was associated with a higher d'_{vs} than the latter. Incorrect responses were misses (congruent, invalid expectation) or false alarms (incongruent, valid expectation). Here, the latter was associated with a higher d'_{vs} than the former. Thus perceptual sensitivity for the attended task was facilitated by valid (task-irrelevant) expectations for the unattended task.

4.4 Discussion

In this paper we have shown that the facilitatory effects of prior expectation on perceptual decision also manifest their influence in metacognitive judgments. We developed a target detection paradigm in which the probability of target presence was manipulated block-wise. This probability, of which participants were informed, significantly biased decision thresholds in the expectation-congruent direction, while leaving sensitivity *d'* unaffected (as ensured by our staircase procedure). In this way we avoided confounding increased type 2 sensitivity with increased type 1 sensitivity (Lau & Passingham, 2006), and were able to assess metacognition, indexed by both type 2 *D'* and meta-*d'*, as a function of perceptual decision and prior expectation. Our main finding was that metacognitive sensitivity increased for expectation-congruent as compared to expectation-incongruent perceptual decisions. Metacognitive sensitivity is determined according to the trial-by-trial correspondence between confidence and accuracy. Importantly, because we offered no trial-by-trial information about the probability of target occurrence, our results cannot be attributed to a trivial relationship between an expectancy cue and the subsequent report. Rather, we found a shift in type 1 threshold with expectation, and a liberalisation of type 2 threshold following an expectation-congruent response to an attended target. This suggests that basing decisions on prior expectations induced a superior placement of type 1 and 2 thresholds for metacognition.

Our effect of expectation on confidence required attention, consistent with some previous work in type 1 tasks (Chennu et al., 2013; Hsu et al., 2014; Jiang et al., 2013; but also see. Kok, Jehee, & de Lange, 2012). However, attention was not required for expectations to shape metacognitive accuracy, and showed no main effect on metacognition either (though under diverted attention, metacognition differed as a function of report). Though perhaps counter-intuitive, this invariance of metacognition to attention is broadly consistent with recent work showing that metacognition is preserved for visual sensory memory, which does not require attention (Vandenbroucke et al., 2014). It is also consistent with research demonstrating above-chance metacognitive accuracy for unattended and unseen target stimuli (Kanai et al., 2010).

4.4.1 Measuring metacognition

To assess how metacognition is affected by expectation we used the type 2 signal detection theory (SDT) measure *D′*. However, the type 2 SDT model underlying *D'* assumes that the probability of making a correct or an incorrect response can be modelled as Gaussian distributions over a type 2 decision axis. This formulation is problematic because such distributions are usually impossible to achieve (Evans & Azzopardi, 2007; Galvin et al., 2003). This issue means that *D'* will not be invariant to type 1 or type 2 criterion shifts (Barrett et al., 2013; Evans & Azzopardi, 2007). In the present study, expectation induced both type 1 and type 2 criterion shifts. As a result, we cannot distinguish between two possible reasons for why *D'* may have increased for expectation-congruent responses. One possibility is that expectation increased the quantity of information available for the type 2 judgment (metacognitive efficacy, Fleming & Lau, 2014). Alternatively, the increase in *D'* could have been driven by a change in criteria placement that indirectly optimised metacognitive sensitivity.

We have modelled and interpreted the results in terms of the latter. Specifically, our model predicts that expectations change the evidence distributions and the criteria shift along the decision axis accordingly. The liberalisation of confidence threshold by expectation, though a source of bias in the numerical value *D'* will take, can be interpreted as reflecting the strength of the perceptual experience (Fleming & Lau, 2014). Therefore rather than being unequivocally problematic, type 2 criteria shifts speak to subjective components of perception.

Our finding that expectation increased *D'* was replicated using the measure meta-*d'* (see section 4.3.5.2. Barrett et al., 2013; Maniscalco & Lau, 2012). Meta-*d’* is robust to changes in type 1 and 2 criteria, however response-conditional meta-*d’* – as required by the analyses presented in this paper - is not (Barrett et al., 2013). The invariance is lost because meta-*d’* measures remove bias by taking a weighted average of the (biased) response-conditional measures. Therefore while we replicated our effect using meta-*d',* we remain unable to ascertain whether expectation improves metacognitive *efficacy* or not. Nevertheless, our results under type 2 *D’* and meta-*d’* together provide converging evidence for the facilitatory effect of expectation on metacognition.

4.4.2 Modelling metacognition

The framework of SDT applied to visual perception emphasises the importance of ‘bottom-up’ processing, whereby afferent sensory signals are repeatedly transformed to generate perceptual decisions at both objective (type 1) and subjective (type 2) levels. However, our data add to an increasing body of work which has demonstrated the importance of top-down processes in shaping perceptual decisions (Bar et al., 2006; Gilbert & Li, 2013; Wacongne et al., 2011). Together, these data pose a challenge to bottom-up models of perception and are difficult to reconcile with standard expressions of SDT.

To formally account for these top-down effects within SDT, we developed a type 2 Bayesian signal detection model which models prior expectations by defining decision threshold as the posterior odds of a target being present. This model successfully predicted an increase in type 2 *D’* following expectation-congruent responses. Diverted attention was modelled by increasing internal noise - as recently proposed by Rahnev et al. (2011). This successfully predicted that the influence of expectation on *D'* would be independent of attention.

We recognise that our model did not capture all aspects of the observed data. In particular, the model predicted an improvement in metacognition following a “no” response, but this was not found empirically. This discrepancy is likely to have arisen from influences on metacognition that were not included in our model, such as the incorporation of additional sources of information relevant to perceptual decision. Nonetheless, by accounting for the main effects of (top-down) prior expectations on *D’,* we have demonstrated the scope for formal synthesis between the traditionally ‘bottom-up’ signal detection theory and ‘top-down’ influences characteristic of alternative frameworks like ‘predictive coding’ or the Bayesian brain (Beck et al., 2009; Clark, 2013; Friston, 2009; Hohwy, 2013; Lee & Mumford, 2003).

4.4.3 From SDT to the Bayesian brain

The increasingly influential predictive coding framework views the brain as a Bayesian hypothesis-tester, and explains perceptual decision as an inference about the most likely cause of incoming sensory input (Clark, 2013; Rao & Ballard, 1999; Seth, 2014a). In this view, top-down expectations constrain perceptual decision according to the prior likelihood of that decision. The sensory input remaining unexplained is termed prediction error, and only this percolates upwards in the sensory hierarchy (Friston, 2010; Rao & Ballard, 2004; Spratling, 2008). The eventual perceptual choice will be the perceptual hypothesis with the highest posterior probability. This framework fits comfortably with our novel finding that under dual-task conditions, sensitivity for the attended (visual search) task was increased when participants held valid expectations pertaining to the unattended (Gabor) task: when prior expectations facilitate decision for the unattended Gabor task, additional processing resources should be available for the attended visual search task (Hohwy, 2012).

Certain predictive coding formulations also explicitly model the importance of the reliability (or 'precision') of the bottom-up signal to perception (e.g. Feldman & Friston, 2010). In this paper we have shown that expectations liberalised subjective confidence judgments for attended (i.e. high precision) targets. Previous work has shown that confidence judgments are a function of both sensory evidence and internal noise (Kepecs, Uchida, Zariwala, & Mainen, 2008; Yeung & Summerfield, 2012; Zylberberg et al., 2012, 2014). This relationship has been likened to a *p*-value, which quantifies the evidence for a hypothesis (mean) and scales with the reliability of that evidence (standard error; Kepecs & Mainen, 2012). In fact, such a formulation of confidence is highly compatible with predictive coding. Bringing these together, decisional confidence could be explained in predictive coding terms, where the mean is the posterior probability of a perceptual hypothesis, and the standard error is the precision of the evidence (Feldman & Friston, 2010). Such a conceptualisation of confidence would explain the congruency-attention interaction found in this paper. It is also consistent with work demonstrating that confidence evolves together with the decision variable (De Martino et al., 2013; Fetsch et al., 2014; Kepecs et al., 2008).

The above account may explain the construction of confidence judgments within a single level of the perceptual hierarchy. However, successful metacognitive evaluations and the subjective aspect of decisional confidence may be a function of uncertainty estimates over multiple hierarchical levels. We leave the theoretical and neural underpinnings of how expectation modulates metacognition open to future research.

4.5 Conclusions

In summary, we show for the first time that top-down prior expectations can influence metacognition for perceptual decision, illustrating the action of priors on complex cognitive functions. Specifically, we found that perceptual decisions which are congruent with valid perceptual expectations lead to increased metacognitive sensitivity, independently of attentional allocation. This finding motivated the development of an extended Bayesian signal detection theoretic model that incorporates top-down prior expectations, and moreover, formally integrates two previously distinct frameworks for perceptual decision: (top-down) predictive coding and (bottom-up) signal detection theory. Finally, measures of metacognition are often used as an indirect measure of awareness (Kanai et al., 2010; Kunimoto et al., 2001; Seth et al., 2008). Therefore, by demonstrating increased metacognitive sensitivity for expected perceptual events, we provide evidence for the existence of a mechanism, based on prior expectations, that underpins metacognitive sensitivity and contributes to our understanding of the brain basis of visual awareness.

5

Rhythmic Influence of Priors in the Phase of Ongoing Occipital Alpha Oscillations

Prior expectations have a powerful influence on perception, biasing both decision and confidence. However, how this occurs at the neural level remains unclear. It has been suggested that spontaneous alpha-band neural oscillations represent rhythms of the perceptual system that periodically modulate perceptual judgements. We hypothesised that these oscillations instantiate the effects of expectations. While collecting scalp EEG, participants performed a detection task that orthogonally manipulated perceptual expectations and attention. Trial-by-trial retrospective confidence judgements were also collected. Results showed that independently of attention, pre-stimulus occipital alpha phase predicted the weighting of expectations on yes/no decisions. Moreover, phase predicted the influence of expectations on confidence. Thus, expectations periodically bias objective and subjective perceptual decision-making together, prior to stimulus onset. Our results suggest that alpha-band neural oscillations periodically transmit prior evidence to visual cortex, changing the baseline from which evidence accumulation begins. In turn, our results inform accounts of how expectations shape early visual processing.

5.1 Introduction

Perception is subject to powerful top-down influences. For example, a highly ambiguous figure can be easily identified following brief priming of object identity (Porter, 1954). Many believe that the feed-forward sensory input is shaped by feedback or recurrent connections from high-level cortical areas to lower-level regions (Gilbert & Li, 2013; Gilbert & Sigman, 2007; Lee, 2002) following a first pass up the sensory hierarchy (Bar, 2003). However, the neuronal mechanisms that integrate top-down and bottom-up signals remain largely unknown (Bar, 2003)

Top-down influences, including priming, context effects and prior exposure, can be parsimoniously construed as a process that biases perceptual inference towards a plausible solution. In line with this, there has been renewed interest in framing top-down influences in terms of probabilistic prior beliefs, or 'expectations' (Summerfield & de Lange, 2014) which, behaviorally, bias perceptual choice (see Chapter 4, or de Lange et al. 2013). It is suggested that expectations are represented in high-level cortical regions prior to the perceptual event, and entrain task-relevant neurons at lower levels to increase sensitivity (Engel et al., 2001). Spontaneous neural oscillations are therefore a promising candidate mechanism for how expectations shape perception.

Oscillations in the alpha range are particularly relevant when considering how expectations influence perception. Theoretical models have associated top-down processes with oscillations in the 8 to 14Hz range (Bastos et al., 2012; Friston, Bastos, Pinotsis, & Litvak, 2014) and recent neurophysiological findings suggest that occipital alpha oscillations primarily propagate in a top-down fashion (van Kerkoerle et al., 2014), supporting the notion that alpha power is intimately related to top-down control (Klimesch et al., 2007; Mathewson et al., 2012; Palva & Palva, 2007). Recent work has revealed that the phase (in addition to power) of pre-stimulus alpha oscillations also predicts various components of perception. These include spatial attention (Busch & VanRullen, 2010), saccadic reaction speed (Drewes & VanRullen, 2011), and perceptual awareness ratings (Mathewson, Gratton, Fabiani, Beck, & Ro, 2009). This has been interpreted as reflecting cycles in the 'preparedness' of the perceptual system (Vanrullen, Busch, Drewes, & Dubois, 2011). In Bayesian terms, prior beliefs (i.e. expectations) are available before stimulus onset. Accordingly, we hypothesised that this ‘preparedness’ should be modulated by expectations: anticipating a perceptual event should bias perceptual inference towards that event. This was tested by asking whether the extent to which decisions are biased by expectation oscillates with pre-stimulus occipital alpha phase.

Perceptual decisions are additionally accompanied by a subjective degree of confidence, which represents belief in one’s decision accuracy and may arise from uncertainty about external (i.e. sensory) or internal noise. Recent work has shown that the decision variable and decision confidence may be encoded together (Kiani & Shadlen, 2009), and arise from the same sensory evidence (Fetsch, Kiani, Newsome, & Shadlen, 2014). In addition to expectations biasing decision, expected perceptual events are associated with greater subjective confidence (see Chapter 4). Following these findings, we additionally hypothesised that pre-stimulus alpha phase would predict the influence of perceptual priors on confidence.

These two hypotheses were tested by adopting a dual-task Gabor detection paradigm which manipulated prior expectations while controlling for the (often conflated) influence of attention (Feldman and Friston 2010; Summerfield and Egner 2009). Prior expectations of target presence or absence were induced by manipulating (block-wise) the probability of Gabor appearance, presented at a contrast that yielded 70% accuracy. The probability was either 25%, such that absence was expected, or 75%, such that presence was expected. A concurrent visual search task diverted attention from the Gabor task in half of the blocks. Critically, the visual search array and Gabor were presented simultaneously following a jittered inter-stimulus interval (ISI; figure 5.1). This allowed us to time-lock our EEG analysis to both Gabor present and Gabor absent trials, and compute independent measures of decision threshold (bias) and detection sensitivity as a function of condition and pre-stimulus EEG phase.

Our first hypothesis was that pre-stimulus alpha phase would predict the extent to which decision threshold is biased by expectation. This would be shown if (1) decision threshold oscillates with pre-stimulus phase and (2) there is some phase angle that predicts ‘yes’ responses when expecting target presence (the 75% condition) while predicting ‘no’ responses when expecting target absence (the 25% condition).

Our second hypothesis was that pre-stimulus alpha phase would also predict expectancy effects on subjective confidence. This would be shown if (1) confidence oscillates with pre-stimulus phase and (2) the same phase that predicts high confidence when expectations are met will predict low confidence when expectations are violated.

5.2 Methods

5.2.1 Participants

Participants were 20 English-speaking subjects (13 female) aged between 20 and 32 years (*M* = 25.6 *SD* = 3.3) with normal or corrected-to-normal vision. One participant’s data were excluded from analysis for being excessively noisy, and a second for having too few trials (<500 vs. mean of 1,100). This was due to excessively slow responding. This left 18 participants’ data for analysis. All participants gave informed, written consent and were reimbursed at £10.30/hour. On average, each session lasted 2.5 hours and two sessions were completed 24 hours apart. Ethical approval was awarded by the University of Sussex ethics committee (C-REC).

5.2.2 Stimuli and Design

The experiment was presented on a 21-inch CRT monitor (100Hz, 1048 x 700 resolution) using Psychtoolbox for Matlab. The experiment was composed of two concurrent tasks: detection of a peripheral Gabor patch and a visual search task in the center of the screen (figure 5.1).

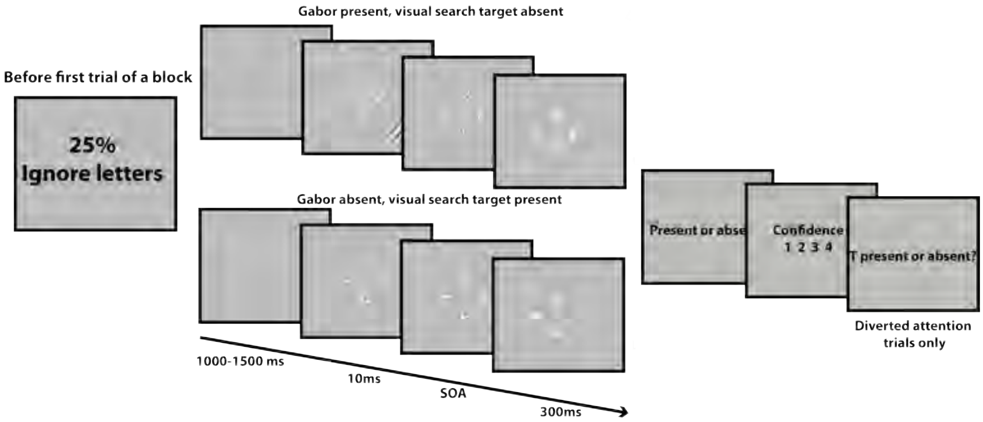

Figure 5.1. Trial sequence.

Before the first trial of a block participants are informed of the experimental condition they are in. “25%” means that the participant is in the 25% of Gabor presence condition and “ignore letters” means that the participant should ignore the visual search array (i.e., they are in the full attention condition). During the trial, a target Gabor is either present (top) or absent (bottom). Similarly, a visual search target T is either present (bottom) or absent (top). Response prompt followed the offset of the masking array.

Trials began with the onset of a white fixation cross. After a jittered inter-stimulus interval (ISI; 1000 to 1500ms) the visual search array appeared. This consisted of four rotated (random orientation of 0° to 359°) white, capitalised letters arranged around fixation (1.43° x 1.43°) at 0°, 90°, 180° and 270°. On 50% of trials the visual search target was absent and all letters were 'L's. On the other 50%, a target ‘T’ replaced one randomly designated ‘L’. To ensure that the task was sufficiently difficult to divert attention from the Gabor task, this array was backwards masked by an array of 'F's. The stimulus onset asynchrony (SOA) between the visual search and masking array was titrated for each individual to equated detection performance to 78% across participants (see Staircases).

On Gabor ‘target present’ trials, a peripheral (3.85° x 4.10° visual angle) Gabor patch (SD 0.89°, *sf* 0.08c/°, phase 45°) was presented in the lower-right quadrant of the screen. On these trials the Gabor and the visual search array appeared simultaneously. The Gabor was presented for 10ms at the contrast resulting in a 70% hit rate (see Staircases).

Following the offset of the visual search array a series of response prompts appeared. Using a key-press, participants made un-speeded judgments of first, Gabor presence or absence, second, confidence that they were correct on an interval scale from 1(no confidence) to 4 (total confidence), and finally, the presence or absence of a ‘T’ in the visual search array.

The experiment had four conditions, constructed in a blocked attention (full, diverted) by expectation (expect Gabor presence, expect Gabor absence) design. Under full attention participants fixated centrally but did not perform the visual search task, thereby allocating full attention to Gabor detection (visual search responses were not requested). Under diverted attention participants performed both tasks, prioritising visual search. Expectation was manipulated by informing participants of the true probability of Gabor presence (as well as the attention condition) before each block began. This was either 25% (expect absence) or 75% (expect presence). After each experimental trial a condition-specific 2 down 1 up staircase titrated the contrast of the Gabor to maintain a consistent hit rate during the long experimental sessions. Expectation-specific staircases controlled for potentially greater levels of sensory adaptation to the Gabor in the 75% condition.

Each block consisted of 12 trials from one of the conditions and blocks were completed in sets of 8 (2 of each condition, 96 trials). Blocks were fully counterbalanced. Participants completed as many blocks as possible in each testing period (always equal numbers of each condition; 6 to 18 runs of each condition per session, M = 11.5). Across participants there was considerable variation in total trials completed due to the cumulative effect of reaction time differences.

After explaining the task to participants they completed a set of practice trials. Next, they completed three staircase procedures (see Staircases) and finally, the experimental trials. Participants were encouraged to take regular breaks and were offered to leave the session early if they became too tired to continue.

5.2.3 Staircases

Following a set of practice trials, participants completed 3 interleaved 2 down 1 up psychophysical adaptive staircase procedures with 8 reversals in order to equate task difficulty across conditions and participants. The visual display was always the same as that in the experimental trials but the instructions and response prompts differed. In the first staircase participants performed Gabor detection while ignoring the visual search array (full attention). Only Gabor present/absent responses were collected. Gabor contrast was titrated to achieve a 70% hit rate (contrast cannot be titrated in target absent trials) under full attention. In the second staircase (3 down 1 up) the Gabor was ignored and participants performed only visual search. Here, only responses to the visual search target were collected (T present/absent). The SOA between the visual search array and the masking array was titrated to achieve 78% accuracy in the visual search task. In the third staircase participants performed both Gabor detection and visual search simultaneously, prioritising visual search and reported both Gabor presence/absence and T presence/absence. Here, Gabor contrast was titrated to achieve a 70% hit rate under diverted attention. The SOA for the visual search display was set to that determined by the second staircase. Confidence judgements were not collected during the staircases.

5.2.4 EEG Acquisition

EEG data were collected on an ANT system at a sample rate of 2048 Hz with no online filtering. Activity was measured continuously from 62 active electrode channels arranged according to the 10/20 system over the scalp. The ground electrode was placed on the forehead and data were averaged across the whole head online. Impedances were kept below 7 kΩ throughout the experimental session. Participants sat in an electrically shielded faraday cage with an external monitor viewed through shielding glass. Their head was stabilised with a chin rest.

5.2.5 EEG pre-processing

EEG data were pre-processed using the EEGLAB toolbox for Matlab. During pre-processing EEG recordings were down-sampled to 256 Hz and high-pass (0.1Hz) filtered with a finite impulse response filter (EEGlab function ‘eegfilt’). EEG data were visually inspected for excessively noisy channels, which were manually interpolated with their two neighbors on a block-wise basis. No participant required more than three channels interpolated (5 participants in total). No interpolated channels were included in analyses presented in this paper. After interpolation data were referenced to participants' average signal. Data were epoched from 1000ms before visual search array (and Gabor target, if present) onset to 500 ms after. Manual artifact rejection was performed on saccade, eye-blink and excessively noisy trials (5% of trials removed on average). For each participant, each electrode and each trial we computed the time-frequency wavelet decomposition of the EEG data. Window lengths of 1 oscillatory cycle at low frequencies (starting at 2Hz) were used. This length linearly increased with frequency band to a maximum of 15 cycles at 50Hz. This decomposition method generated wavelet coefficients for 49 log-spaced frequencies and 242 time points.

5.2.6 Analysis

5.2.6.1 EEG: Electrode region of interest.

We had an a priori hypothesis that top-down influences of prior expectation would be observable over occipital regions. Initial analyses were therefore restricted to the occipital electrodes O1, Oz and O2. Because phase at some time-frequency point will differ across electrodes, analyses were further restricted to one electrode per participant and session. To control for differences in electrode placement, electrode ROIs (eROIs) were determined on a participant-by-participant and session-by-session basis according to their sensitivity to the Gabor detection task. The grand-averaged ERP indicated a negative deflection following hits relative to misses in the 75-200ms range. Each participant’s session-specific eROI was therefore chosen as the occipital electrode (i.e. O1, Oz or O2) that showed the greatest event-related potential (ERP) amplitude, as defined below. To compute the ERPs a 200ms pre-stimulus baseline was subtracted from each epoch. Epochs in which hits (respectively, misses) were made were averaged together. For each response type (hit or miss) we obtained the maximal local peak amplitude (LPA) in the 75ms-200ms period. LPA is defined as the greatest amplitude within a range of time points such that this peak is greater than the average amplitude of the surrounding 7 time points (Luck, 2005). This method minimises the chance of selecting spurious spikes. The eROI for each participant was chosen as the occipital electrode that showed the greatest value for LPA_{hit} – LPA_{miss}. Subsequent analyses on phase were restricted to these eROIs.

5.2.6.2 EEG: Phase opposition analysis.

Next, we sought to determine if, for our eROI, spontaneous EEG phase differed at any time point and in any frequency band between 'reported present' (yes) and 'reported absent' (no) trials. This was done in order to isolate candidate time-frequency regions in which expectation might interact with the influence of EEG phase. The relationship between phase and response was quantified with the measure phase opposition (Vanrullen et al., 2011), which is defined as the mean of phase locking values (PLV) for yes and for no responses. Phase locking value measures the extent to which phase angle at some time-frequency point over one electrode is predicted by either (A) phase at the same time-frequency point over another electrode or (B) a behavioural response (as in the present paper). Here, we used PLV as a measure of the relationship between ongoing phase and response. Because yes and no responses encompass all possible responses and because stimulus onset is unpredictable (randomised ISIs), the joint PLV across all trials is expected to be small (no different from chance). However, if EEG phases for a given behavioural response are clustered about some angle (necessarily different for yes vs. no) then the individual PLVs for both yes and no responses, and therefore the resulting phase-opposition value, will be high (up to 1 for perfect phase-opposition; see Vanrullen, Busch, Drewes, & Dubois, 2011 for additional details). High (and statistically significant) values indicate that phase predicts a yes versus a no response. For a set of n trials where response R is given and where C(R) is the complex coefficients of the wavelet transform, PLV_R and phase opposition PO for responses R_1 and R_2 are defined as follows:

$$PLV_R = \left|\frac{1}{n}\sum_{n}\frac{C(R)}{|C(R)|}\right| \qquad PO_{R_1,R_2} = \frac{PLV_{R_1} + PLV_{R_2}}{2}$$

This measure PO is similar to the phase bifurcation index (PBI; Busch, Dubois, & VanRullen, 2009). PBI is defined as (PLV_{R1} -PLV_{ALL}) X (PLV_{R2} - PLV_{ALL}), that is, the baseline-corrected product of phase locking values for response 1 and for response 2. We preferred the additive measure PO, because PBI can give unreliable results when taking the product over very small values. Moreover, because PO is additive it is robust to differences in trial counts between 'yes' and 'no' trials: any baseline correction applied to empirical PO values would be equally applied to bootstrapped PO values and cancel out.

PO between yes and no responses was separately calculated for each level of attention and expectation. Separate calculation of PO for each level of expectation was necessary because we hypothesised that the phases predicting ‘yes’ (respectively, ‘no’) would differ as a function of expectation. The four PO time-frequency maps corresponding to each experimental condition were averaged together.

At each time-frequency point, PO statistical significance was assessed by estimating the mean and standard deviation of the null distribution from 8000 bootstrapped samples per participant. To obtain bootstrapped samples, responses were pseudo-randomly assigned to trials such that the number of yes and no responses stayed the same. PO was then recalculated. This method removed any relationship between the EEG signal and behaviour. *Z*-scores and *p* values were computed by comparing empirical PO values to the mean and standard deviation of the bootstrapped values. P values were false discovery rate (FDR) corrected for multiple comparisons over all frequencies and all pre-stimulus time-points.

5.2.6.3 EEG: Phase modulation of perceptual decision

The time-frequency representation of phase opposition values revealed that phase is related to the subjects’ response (see above and fig. 5.3B). However, we did not know (and aimed to determine) whether the “optimal” phase for a yes response is comparable for the different expectation conditions. To determine whether the influence of expectation on decision is predicted by pre-stimulus phase in some frequency band, a follow-up analysis was run in which the data were restricted to a time-frequency region of interest. The time-frequency ROI was taken as the point of maximal phase opposition (PO) significance. Critically, there was no circularity in this analysis because PO values had been collapsed across levels of expectation.

For each participant, each condition and each trial, the phase at the time-frequency ROI was binned into one of 6 phase bins. For each bin we then computed within-subject signal detection theoretic (SDT) outcome variables *d'* (sensitivity), *c* (decision threshold/bias) confidence (percentage of trials reported with high confidence). This provided values of each SDT outcome as a function of condition and phase bin. Using 6 bins enabled a sufficient number of trials for SDT estimates to be reliable.

5.2.6.4 Signal detection theoretic (SDT) outcomes.

To obtain separate measures of detection sensitivity and decision bias, we used signal detection theory (SDT, see Chapter 3). For each experimental condition, trials were categorised into hits, misses, false alarms and correct rejections. Hit rate and false alarm rate are then defined as:

$$Hit\ rate = \frac{hits}{hits + misses},$$

$$False\ alarm\ rate = \frac{false\ alarms}{false\ alarms + correct\ rejections}$$

From these quantities, detection sensitivity for the Gabor target, *d',* and decision threshold *c* are given by:

$$d' = Z(Hit\ rate) - Z(false\ alarm\ rate), \quad c = -\frac{Z(Hit\ rate) + Z(false\ alarm\ rate)}{2}$$

where Z is the inverse normal cumulative distribution function. Note that for decision threshold *c*, positive values represent a conservative bias (more likely to report Gabor absence) and negative values represent a liberal bias (more likely to report Gabor presence).

In computing these measures we used the log-linear rule, which adds 0.5 to the total number of hits, misses, false alarms and correct rejections. This ensures SDT outcome variables can be computed for all conditions and phase bins, and also acts as a Bayesian prior on a *d'* of zero.

5.2.6.5 CONFIDENCE.

Confidence ratings were collected on a four-point scale. To account for individual differences in how the scale was used (mean confidence: 2.92, range: 2.34 - 3.47) we collapsed ratings onto a binary scale. This was achieved by calculating each participant's mean confidence across all conditions then categorising each rating as high (greater than the mean) or low (lower than the mean). Note that we did not use a median split because here, the median is always an integer.

5.2.6.6 STATISTICAL ANALYSES

Data were collected over two experimental sessions and collapsed across them. Session number did not significantly interact with any other factors under any behavioral dependent variable. For each participant reported confidence was collapsed onto a binary scale using a mean split (median split of a four-point integer response scale cannot be formed). Analyses were conducted using Matlab, CircStat toolbox for Matlab (Berens, 2009) for circular statistics, and SPSS. Where appropriate, *p* values were FDR (false discovery rate) corrected. Circular statistics were corrected for the binning of phase angles where appropriate. Unless otherwise specified, data subjected to within-subjects ANOVAs met the assumption of sphericity.

5.3 RESULTS

5.3.1 EXPECTATION AND ATTENTION SEPARATELY INFLUENCE CONTRAST SENSITIVITY

To determine the success of our attention manipulation we asked whether diverting attention with the visual search task decreased contrast sensitivity (as determined by the psychophysical staircases). Mean Gabor contrast was subjected to an Attention (full, diverted) x Expectation (25%, 75%) repeated-measures ANOVA. This revealed a significant main effect of Attention, *F*(1,17) = 22.60, *p* < .001, η_p^2 = .57, such that contrast sensitivity was significantly greater (i.e. contrast threshold decreased) in the full (19.8%±1.2%) than diverted (25.7%±1.3%) attention condition. Our manipulation of attention was therefore successful. The ANOVA also revealed a significant main effect of Expectation, *F*(1,17) = 8.50, *p*= .010, η_p^2 = .33, whereby contrast sensitivity was significantly greater in the 75% (22.3%±1.1%) than the 25% (23.3%±1.1%) condition. This is likely to be an outcome of more Gabor exposure in the 75% than the 25% condition, which was controlled by implementing running staircases during the experimental phase (see Staircases). The interaction between Attention and Expectation was not significant *F*(1,17) = 1.26, *p* = .278, η_p^2 = .07. Results are represented in fig. 5.2A.

5.3.2 Expectations bias decision and increase subjective confidence

The main behavioral analyses presented here used Signal Detection theory (for details, see Methods). To ensure that our expectation manipulation successfully biased choice, decision threshold *c* was calculated as a function of condition. Here, *c* > 0 represents a conservative bias (i.e. towards reporting ‘no’) whereas *c* < 0 represents a liberal bias (i.e. towards reporting ‘yes’). An Attention (full, diverted) x Expectation (25%, 75%) repeated-measures ANOVA revealed that *c* was significantly affected by Expectation, *F*(1,17) = 70.33, *p* < .001, η_p^2 = .80. As predicted, *c* was significantly more conservative in the 25% than the 75% condition (M_{diff} = 0.21±0.03, figure 5.2B), meaning that decisions were more biased towards absence in the ‘expect absent’ (25%) than the ‘expect present’ (75%) condition. There was neither a significant main effect of Attention, *F*(1,17) = 0.01, *p* = .952, η_p^2 < .01 nor a significant interaction between factors, *F*(1,17) = 1.45 *p* = .244, η_p^2 = .08.

To determine whether detection sensitivity had been successfully equated across conditions an Attention x Expectation repeated-measures ANOVA under detection sensitivity *d’* was run. This revealed a significant main effect of

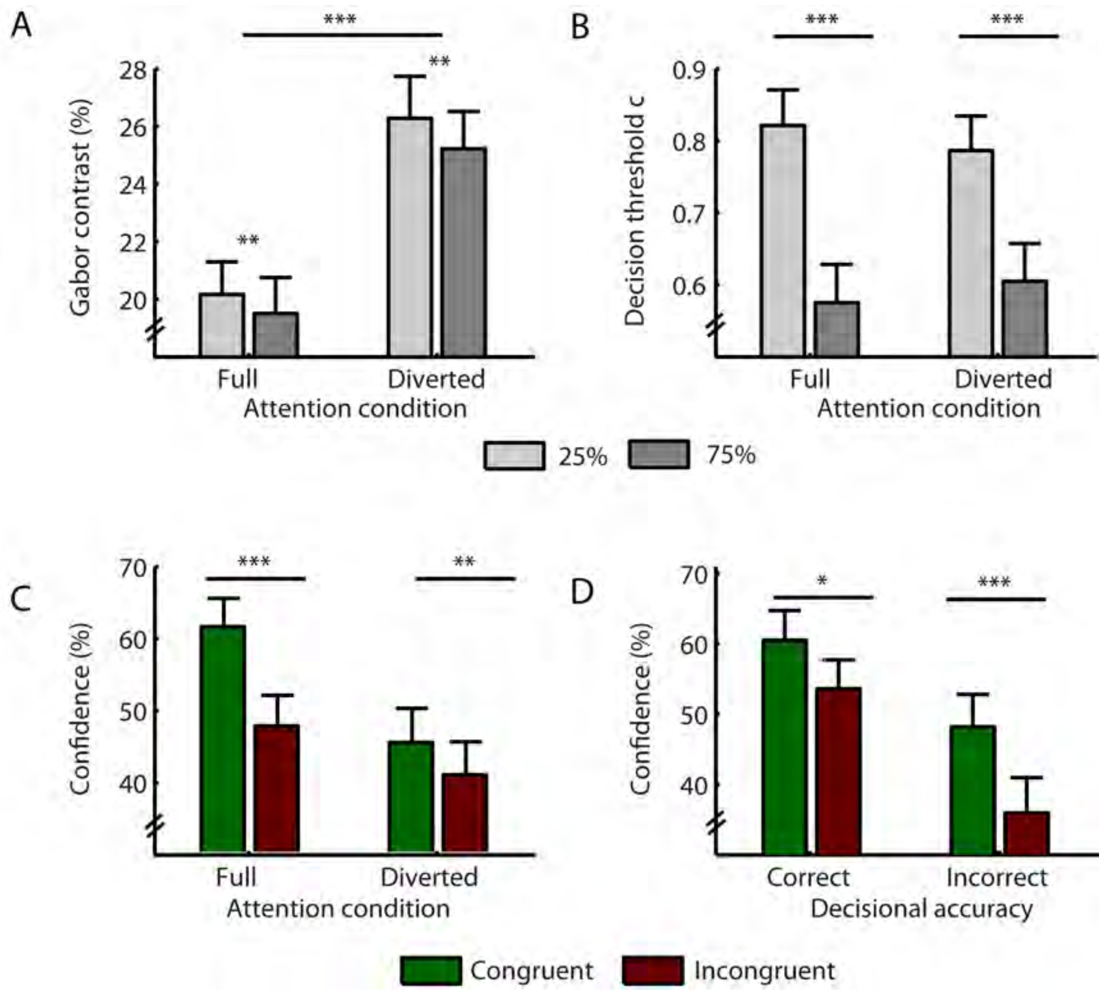

Figure 5.2. Behavioural results.

A Mean contrast at which the Gabor was presented over the course of the experiment in each condition. Significant main effects of both attention and expectation. **B.** Effects of attention and expectation on decision threshold c. Independently of attention, decision threshold in the 25% condition is significantly more biased towards ‘no’ responses than in the 75% condition. **C**. Effects of attention and expectation-report congruence on confidence. Congruent responses are reports of presence/absence in the 75%/25% condition, and vice versa for incongruent responses. Confidence is higher for congruent than incongruent reports in both attention conditions, but the effect of congruence is greater under full attention. The main effects of both attention and congruence are also significant. **D**. Effects of accuracy and expectation-report congruence on confidence. Confidence is higher for congruent than incongruent reports for both correct and incorrect responses, but the effect of congruence is greater in the incorrect case. The main effects of both accuracy and congruence are also significant. Error bars represent within-subject SEM. *p < .05, ** p < .01, *** p ≤ .001.

Expectation, $F(1,17) = 52.85$, $p < .001$, $\eta_p^2 = .76$, such that *d'* was greater in the 25% (2.60±0.09) than the 75% (2.23±0.09) condition. This small difference was an unavoidable consequence of liberalising decision threshold while ensuring a constant hit rate. The main effect of Attention, $F(1,17) = 0.46$, $p = .507$, $\eta_p^2 = .03$, and its interaction with Expectation, $F(1,17)= 0.23$, $p = .655$, $\eta_p^2 = .01$, was not significant.

Chapter 4 showed that expectations increase subjective confidence and improve metacognitive accuracy. On this basis, we hypothesised that pre-stimulus phase would modulate the influence of expectations on confidence. To address this at the behavioral level, the next analyses determined whether this finding was replicated.

In the 25% condition, where Gabor absence is expected, the expectation-congruent report is 'no', whereas in the 75% condition, where Gabor presence is expected, the expectation-congruent report is 'yes'. The reverse defines expectation-incongruent reports. A within-subjects Attention (full, diverted) x Accuracy (correct, incorrect) x Congruence (expectation-congruent, incongruent) repeated-measures ANOVA under confidence was run. Results showed that confidence was higher under full than diverted attention, $F(1,17) = 17.67$, $p = .001$, $\eta_p^2 = .51$, for correct than incorrect responses, $F(1,17) = 42.22$, $p < .001$, $\eta_p^2 = .71$, and for congruent than incongruent decisions, $F(1,17) = 19.07$, $p < .001$, $\eta_p^2 = .53$.

As shown in figure 5.2C, a significant attention x congruence interaction, $F(1,17) = 14,83$, $p = .001$, $\eta_p^2 = .47$, revealed that diverting attention reduced the effect of congruence on confidence ($M_{diff} = 4.6\%$ $SE_{diff} = 1.4\%$) relative to full attention ($M_{diff} = 14.1\%$ $SE_{diff} = 3.2\%$). Congruence still increased confidence in both attention conditions (diverted: $t(17) = 3.25$, bootstrapped $p = .006$; full: $t(17) = 4.41$, bootstrapped $p = .001$).

As shown in figure 5.2D, a significant accuracy x congruence interaction, $F(1,17) = 8.48$, $p = .010$, $\eta_p^2 = .33$, revealed that the influence of congruence on confidence was greater for incorrect ($M_{diff} = 12.0\%$ $SE_{diff} = 2.6\%$) than correct

(M_{diff} = 6.7% SE_{diff} = 2.1%) responses. Crucially, congruence increased confidence in both cases (incorrect: *t*(17) = 4.67, bootstrapped *p* = .001; correct: *t*(17) = 3.29, bootstrapped *p* = .014) indicating that the influence of congruence on confidence is not confounded by differences in decisional accuracy.

No other significant effects were found (attention x accuracy, *p* = .102, η_p^2 = .15; attention x accuracy x congruence, *p* = .975, η_p^2 < .01). Thus, effects under confidence reported in Chapter 4 were replicated: expectations liberalise confidence, and the effect was weaker (but present) under diverted than full attention.

Are these changes in confidence associated with changes in metacognitive bias? To address this question we estimated each participant's response-specific confidence thresholds τ_+ and τ_- for each level of attention and expectation (see section 3.6.1). These were scaled by *d'* and subtracted from type 1 *c.* That is, we computed the measures

$$C'_{1-} = \frac{\tau_-}{d'_-} \quad C'_{1+} = \frac{\tau_+}{d'_+}$$

These tell us how far the confidence thresholds extend from the decision threshold. The further away they are, the more evidence the participants needs to assign high confidence to their choice (i.e. the more *conservative* their threshold). This measure of metacognitive bias was log-transformed (because raw values are bounded by zero) and subjected to an Attention by Expectation by Report repeated-measures ANOVA. Independently of attention, report and expectation interacted, *F*(1,17) = 12.17, *p* = .003, η_p^2 = .417 (figure 5.3).

Follow-up bootstrapped *t*-tests revealed that for "no" reports, confidence threshold C'_{1-} was significantly closer to *c* (i.e. more liberal) in the 25% (expect absent) than the 75% (expect present) condition, *p* = .004. However a congruence effect was not found for "yes" reports (*p* = .237). This may be due to a presence of floor effects: confidence for "yes" reports was already very liberal, probably because the Gabor target had an abrupt onset, leading to visual pop-out'.

This analysis shows that confidence thresholds for reports of target absence were liberalised when absence was expected, such that less evidence was required to report the choice with high certainty. In turn, this means that expectations don’t only shape confidence indirectly (by shifting decision bias), but they also may target metacognitive thresholds.

Finally, we determined whether expectations improved metacognitive accuracy as defined by type 2 D and meta-*d’*/*d’*. As in chapter 4, each measure of metacognition was subjected to separate Attention x Expectation x Report repeated-measures ANOVAs. Under both type 2 D’ and meta-*d’*/*d’* a significant

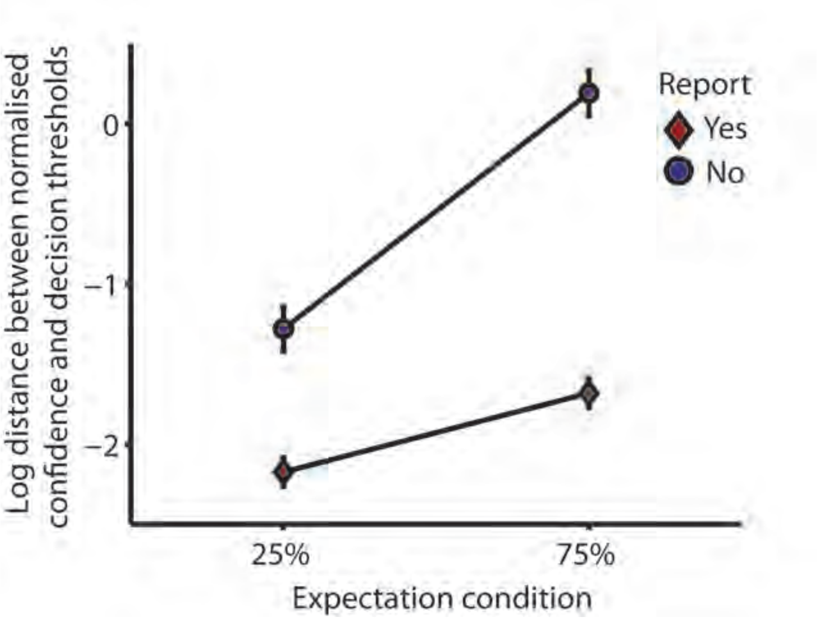

Figure 5.3 Metacognitive bias as a function of expectation and report.

This panel depicts the log distance between type 1 criterion and type 2 confidence thresholds. Larger values mean that more evidence is needed to report ‘confident’. When participants reported “no” (blue circles), confidence was more liberal, i.e. took a lower value, in the 25% than the 75% condition. That is, when expecting target absence, perceived absence required less evidence to be reported with high confidence. For “yes” responses (red diamonds) there is no difference in thresholds between the 25% and 75% conditions, however this appears to be driven by floor effects. Error bars represent within-subjects SEM.

3-way interaction was found (type 2 D’: *F*(1,17) = 2.25, *p* = .025, η_p^2 = .261. Meta-d’/d’: *F*(1,17) = 6.20, *p* = .023, η_p^2 = .267).

Expectation-response congruence influenced metacognition under full attention (type 2 D’: *F*(1,17) = 6.74, *p* = .019, η_p^2 = .284. Meta-*d’*/*d’*: *F*(1,17) = 9.50, *p* = .007, η_p^2 = .358) but not diverted attention (type 2 D’: *F*(1,17) = 3.92, *p* = .540, η_p^2 = .023. Meta-d’/d’: *F*(1,17) = 1.54, *p* = .700, η_p^2 = .009). Thus, we replicated the finding of Chapter 4 that expectations improved metacognitive accuracy.

In summary, our paradigm successfully manipulated attention and expectation: contrast sensitivity increased in the presence of full attention, and expectation biased perceptual decisions. There was a small difference in *d*’ across levels of expectation but not across levels of attention. Expectation further increased confidence, such that participants were more confident in their Gabor detection reports when that report had been congruent with their prior expectations. For “no” reports, this was driven by changes in the threshold for reporting “confident”.

While these effects of expectation were present at the behavioral level, they are not necessarily modulated by pre-stimulus brain oscillations. The next analyses first determined whether oscillatory phase predicts perceptual decision irrespective of expectation, and then determined whether the predictive value of oscillatory phase reflects prior expectations.

5.3.3 Perceptual decision is predicted by occipital alpha phase

Before addressing the question of whether the effect of expectation on decision is modulated by pre-stimulus phase over visual regions, we checked that pre-stimulus phase predicted perceptual choice, irrespective of expectation. Analyses were restricted to the occipital electrode (O1, Oz or O2) that showed the greatest post-stimulus response to the Gabor task. This method gave, for each participant and for each of the 2 sessions, a single electrode (eROI) that was involved in early post-stimulus processing. eROIs were extracted by selecting the occipital electrode with the greatest event-related potential (ERP) amplitude for hit relative to miss trials (M_{diff} = 0.75µv, SD_{diff} = 0.64µv, see Methods for details).

The predictive value of phase in perceptual decision was assessed using the measure phase opposition (PO). PO is the average of phase-locking values (PLV) for two responses – here, yes and no (Vanrullen et al., 2011) - and therefore reflects the extent to which pre-stimulus phase predicts subsequent choice (see Methods for details). For response R and complex wavelet coefficients C, PLV and PO are defined as:

$$PLV_R = \left|\frac{1}{n}\sum_n \frac{C(R)}{|C(R)|}\right| \qquad PO_{R_1,R_2} = \frac{PLV_{R_1} + PLV_{R_2}}{2}$$

PO values for each time-frequency point were calculated separately for each level of attention and expectation and subsequently collapsed across expectation conditions. This was done because for this initial analysis we were seeking time-frequency regions in which EEG phase predicted decision, but not explicitly seeking time-frequency regions in which the influence of phase depended on expectation. Averaging over conditions means phase effects are still detectable if expectation changes (or even reverses) the preferred phase for yes or no responses. Interactions between phase and expectation were run in a separate follow-up analysis, thereby avoiding ‘double-dipping’.

To obtain *p*-values, PO values were compared to the null distribution by pseudo-randomly allocating a behavioral response to each phase angle at each time-frequency point. This process was repeated for each session and each condition 2000 times (8000 in total), giving 1.8 x 10^{70} bootstrapped samples over all participants. The *p*-values were FDR-corrected over the entire pre-stimulus region (-1000ms to stimulus onset) and over all frequencies.

This analysis revealed a region of significant phase opposition in the pre-stimulus alpha range over all trials, which reached maximum significance at 10Hz, 119ms prior to stimulus onset, ($p = 10^{-7}$, $\alpha_{FDR} = 10^{-2.6}$, figure 5.4A,left). This means that pre-stimulus occipital alpha phase predicts yes versus no responses. Given that phase-modulation of perceptual hit rate has been shown to be dependent on attention (Busch & VanRullen, 2010), we then split phase opposition values into two separate maps, one for each level of attention. Significant phase opposition was present under full attention ($p_{-119ms,\ 10Hz} = 10^{-4}$, $\alpha_{FDR} = 10^{-2}$, figure 5.4A, center), and was indeed reduced in extent (but present) under diverted attention ($p_{-119ms,\ 10Hz} = 10^{-5}$, $\alpha_{FDR} = 10^{-3}$, figure 5.4A, right), consistent with previous work.

This result shows that pre-stimulus occipital alpha phase predicted decision, but we do not yet know whether decision bias or detection sensitivity was fluctuating. This question was addressed in the next section.

5.3.4 Pre-stimulus occipital alpha phase predicts decision thresholds

Previous studies on pre-stimulus phase have not been able to separate sensitivity from decision bias because phase analyses have only time-locked to target-present trials. Whereas target-absent trials usually have no obvious reference point for the phase analysis (when using a randomised inter-trial interval), here the onset of the search array served as a reference point for both Gabor-present and Gabor-absent phase determination. This allowed us to calculate the theoretically independent measures *c* (decision threshold) and *d’* (detection sensitivity).

Computing these values required binning phase angles from each trial. We needed data from just one time point, because pooling phase angles over time points results in associating multiple, systematically rotating phase angles with a single behavioral response. Similarly, phase angles from differing frequency bands cannot be compared in terms of their position in an oscillation. We extracted phase angles from each epoch from the eROIs at the -119ms, 10Hz time-frequency point: the point of maximal PO significance. Each phase angle was then binned into one of 6 phase bins.

By considering responses on those trials this gave, for each participant, an associated set of hits, misses, false alarms and correct rejections as a function

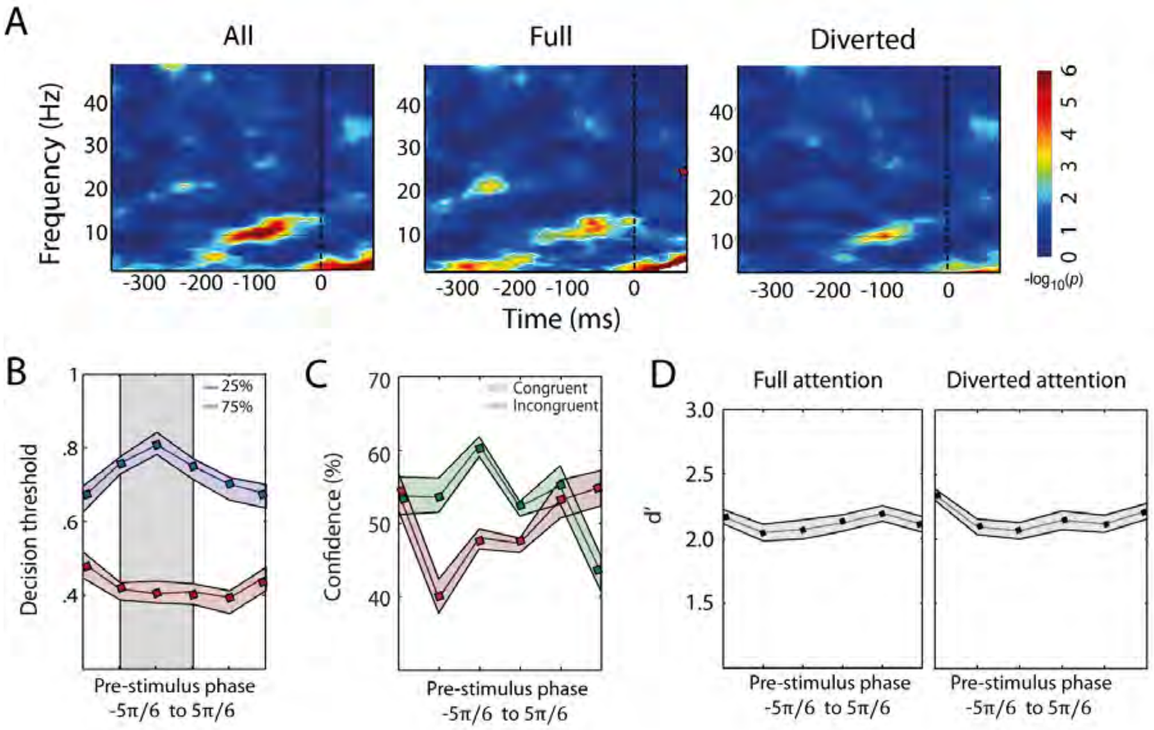

Figure 5.4. EEG results.

A. Time-frequency representation of phase opposition between yes and no reports over the eROI for (left) all trials, (middle) full attention, and (right) diverted attention. The vertical dashed line represents stimulus onset. The colour scale represents log-transformed p-values. Regions that survive FDR correction are outlined in white.

B. Relationship between decision threshold c and binned occipital 10Hz phase at -119ms. The blue phase-criterion function represents results from the 25% (expect absent) condition and the red phase-criterion function represents results from the 75% (expect present) condition. Grey shading indicates the phase values which maximally predict the influence of expectation on decision: decisions are maximally biased towards reporting ‘no’ in the expect 25% condition, but towards ‘yes’ in the 75% condition. Shaded outlines represent within-subjects SEM

C. Relationship between confidence and pre-stimulus 10Hz phase at -119ms. Congruent responses are reports of presence/absence in the 75%/25% conditions and vice versa for incongruent responses. Confidence significantly fluctuates with phase for both congruent (green) and incongruent (red) reports. Shaded regions represent within-subjects SEM.

D. Relationship between detection sensitivity d' and pre-stimulus 10Hz phase at -119ms for the full (left) and diverted attention (right) conditions. Sensitivity does not fluctuate with phase in either condition. Shaded regions represent within-subjects SEM.

of phase bin. Trials were further categorised according to experimental condition. In turn, for each participant we could calculate *d'* and *c* as a function of phase bin, attention and expectation. Note that in splitting trials according to bin, the resulting six values of *c* per condition will not average exactly to the single value of *c* per condition when computed irrespective of phase bin.

First, we asked whether pre-stimulus phase predicts decision threshold by running an Attention (full, diverted) x Expectation (25%, 75%) x Phase bin (1 to 6) repeated-measures ANOVA on decision threshold *c*. Only interactions with phase bin are reported. This analysis revealed no significant main effect of Phase, $F(5,85) = 0.66$, $p = .670$, $\eta_p^2 = .04$, no significant Attention by Phase bin interaction, $F(5,85) = 0.38$, $p = .862$, $\eta_p^2 = .02$, and no significant three-way interaction, $F(5,85) = 0.66$, $p = .650$, $\eta_p^2 = .04$. Critically, there was a significant two-way interaction between Expectation and Phase bin, $F(5,85) = 2.64$, $p = .029$, $\eta_p^2 = .13$. This interaction is depicted in figure 5.4B, and is such that, as hypothesised, (1) *c* appears to oscillate with phase in both expectation conditions and (2) the two phase-criterion functions appear to be in anti-phase.

These curves being in anti-phase mean that the range of phase values related to highest *c* in the 25% condition (conservative, expectation-congruent) is similar to the minimum values for *c* in the 75% condition (liberal, expectation-congruent).

This range is consistent with what we would expect from the optimal phase for perceptual priors to influence perceptual decision. At π rad away from this range, phase predicted the most liberal responses in the 25% condition (incongruent) and the most conservative responses in the 75% condition (incongruent). This suggests that in *this* range of phase the top-down priors exert their weakest influence, and that the relative effect of perceptual priors is minimal. We assume that here, the influence of bottom-up signals is therefore maximal.

Supporting part of our first hypothesis, this indicates that independently of attention, the extent to which pre-stimulus occipital alpha phase predicted decision threshold differed in the 25% (expect absent) and 75% (expect present) conditions.

Figure 5.4B suggests that *c* oscillates in both conditions (both functions are sinusoids), but that the same phases predict opposing responses (the functions are in anti-phase). However, we have not yet determined this statistically. This was the aim of our next two analyses.

5.3.5 Prior expectations change the response predicted by pre-stimulus alpha phase

Does phase predict *c* in both expectation conditions? To check whether the phase-criterion functions were sinusoids we tested whether the distance between the peak and trough of each function was π rad. We used a circular *v*-test, which tests the hypothesis that a set of angles (here, the peak-to-trough distance) is significantly clustered about some specified angle (here, π rad). This analysis revealed that indeed, the peak-to-trough distance was approximately π rad in both the 25% (v = 43.98, p < .001) and the 75% (v = 12.56, p = .044) conditions. This means that both functions are sinusoids, and therefore that phase predicts criterion in both the 25% and 75% conditions.

Next we asked whether the two phase-criterion functions were in anti-phase. This was the final, key step in testing whether expectations were reflected in pre-stimulus phase. A circular *v*-test, testing whether the peak-to-peak difference between the two phase-criterion functions was significantly clustered about π rad, revealed this to be the case, v = 43.98, p < .001. Thus, the two functions are in anti-phase, and the same phases that predict ‘yes’ when expecting target presence predict ‘no’ when expecting target absence. These phases are therefore those at which expectations exert their greatest effect on decision.

In summary, we have supported our first hypothesis: that the influence of expectations on decision is oscillating with pre-stimulus alpha phase. We do not claim that a decision threshold is set at or before stimulus onset, because clearly, sensory evidence is not yet available to the visual system. Rather, our data show that prior to stimulus onset, ongoing alpha phase biased the position of a decision threshold that is set later in time.

5.3.6 Rhythmic fluctuations in confidence

Our second hypothesis was that pre-stimulus alpha phase would also predict the influence of expectations on confidence. Behaviorally, confidence increases for expected percepts. Consistent with this, our behavioral analyses showed that confidence for expectation-congruent reports (i.e. reporting 'yes' in the 75% condition or reporting 'no' in the 25% condition) was higher than for incongruent reports (i.e. reporting 'no' in the 75% condition or reporting 'yes' in the 75% condition). Therefore, if phase predicts the influence of expectations on confidence then there should be a range of phase angles which predict high confidence when congruent reports were made, but low confidence when incongruent reports were made. This set of phases would be the optimal phases for expectations to shape confidence.

The four-point scale was collapsed into a binary confident/guess reports by performing a mean split on individual participants' reports. Next, we computed participants’ percentage of decisions reported with high confidence, as a function of phase bin, attention, and expectation-response congruence.

An Attention x Congruence x Phase bin repeated-measures ANOVA under confidence revealed a significant main effect of phase bin (p < .001), but the phase-confidence function was not sinusoidal and therefore does not reflect the existence of an optimal phase for high confidence. The three-way interaction was also non-significant (p = .198, η_p^2 = .08). Crucially, the analysis did reveal a significant 2-way Congruence x Phase bin interaction, F(5,85) = 4.10, p = .002, η_p^2 =.19.

To break down this interaction we tested whether confidence oscillated with phase at either level of congruence. As in the analysis under decision threshold, circular v-tests tested the peak-to-trough difference of the two phase-confidence functions against π. These revealed that subjective confidence oscillated with pre-stimulus alpha phase for both expectation-incongruent, *v* = 34.56, *p* < .0001 and expectation-congruent, *v* = 25.13, *p* < .001, responses (figure 5.4C).

As was the case for the decision threshold analysis, visual inspection of the figure suggests that the two functions are in anti-phase: phases associated with relatively high confidence for congruent reports are associated with relatively low confidence for incongruent reports. This was confirmed statistically with a circular *v*-test that showed the peak-to-peak distance between the two phase-confidence functions to be significantly clustered about π rad, *v* = 43.98, p < .0001. In turn, this analysis indicates that the two functions are in anti-phase.

Interestingly, the phase at which congruent yes/no responses are most likely appears similar to that at which congruence maximally predicts confidence (see figure 5.4C and 5.4B, respectively): the peak of the phase-expectation function (the 25% minus the 75% sinusoid) appears associated with high confidence for congruent reports, but low confidence for incongruent reports.

In summary, our results suggest that at phases where prior expectations exerted stronger influences on decision; confidence was high for the expectation congruent report, but low for expectation-incongruent reports. This means that when the influence of priors was strong, confidence increased for predicted perceptual events, but decreased when expectations were violated. Together with the results under decision threshold, these data suggest a 10Hz alternation in the extent to which perceptual priors bias both objective and subjective decision-making.

5.3.7 Alpha phase does not predict perceptual sensitivity

Confidence is typically correlated with accuracy, such that participants are more confident when they are correct than when they are incorrect. Previous work has implicated pre-stimulus alpha phase in the detection of perceptual stimuli (Dugué, Marque, & VanRullen, 2011; Mathewson et al., 2012; Rohenkohl & Nobre, 2011), however previous studies have not been able to time-lock the phase analysis to target-absent as well as target-present trials. In turn, it is unclear whether these results reflect alternations in decision biases or in perceptual sensitivity. If sensitivity is predicted by pre-stimulus alpha phase, our results under confidence may simply reflect fluctuations in *d'*.

Our results under *c* implicate alpha phase in decisional biases, however to ascertain whether alpha phase is also implicated in sensitivity we ran an Attention x Expectation x Phase bin rmANOVA under *d'.* This revealed no significant main effect of Phase bin, *F*(5,85) = 1.65, *p* = .156, η_p^2 = .09, nor any significant interactions (Attention x Phase: *F*(5,85) = 0.86, *p* = .507, η_p^2 = .05 (figure 5.4D); Expectation x Phase *F*(5,85) = 0.37, *p* = .868, η_p^2 = .02; Attention x Expectation x Phase, *F*(5,85) = 0.88, *p* = .499, η_p^2 = .05).

An analogous Bayesian repeated-measures Attention x Expectation x Phase bin ANOVA was run on JASP using a Cauchy prior of 0.8 HWHM. This revealed evidence for the null hypothesis of no main predictive effect of phase (BF = 0.025), as well as no predictive effect of phase that depended on attention (BF = 0.003), expectation (BF = 0.001) or both attention and expectation (BF < .0001).

Previous studies have found that it was useful to realign each participant's phase-hit rate function in order to correct for individual differences in optimal phases for perceptual sensitivity (Busch & VanRullen, 2010). Even using this method, however, we found no evidence for phase predicting *d'* under either full (*p* = .787) or diverted (*p* = .407) attention.

Together, these data robustly show that pre-stimulus alpha phase does not predict detection sensitivity. Rather, the data support the interpretation that alpha phase reflects fluctuations in objective and subjective decisional biases.

5.4 Discussion

The present experiment implemented a paradigm that both separated the influences of expectation from those of attention, and allowed pre-stimulus oscillations to be time-locked to both target-absent and -present trials. Critically, this design enabled us to compute signal detection theoretic measures as a function of phase and condition, and in turn separate phase-modulation of detection sensitivity from phase-modulation of decision threshold.

Our results show that top-down expectations rhythmically bias perceptual decision-making in the pre-stimulus period, such that the extent to which expectations biased decision was predicted by the phase of pre-stimulus occipital alpha oscillations. The data revealed that decision threshold was predicted by phase both when expecting target presence and when expecting target absence. However, expectation flipped the relationship between phase and criterion (decision threshold), that is, the phase-criterion functions were in anti-phase: the same phases that predicted biases towards reporting ‘no’ when expecting target absence predicted biases towards reporting ‘yes’ when expecting target presence. These phases correspond to the optimal phases for expectations to influence perception.

Importantly, we do not claim that perceptual priors entrained alpha oscillations, as is the case for temporal predictions (e.g. Rohenkohl and Nobre 2011; Samaha et al. 2015). Rather, priors determined whether a specific phase angle facilitated a ‘yes’ or a ‘no’ judgment. This effect of pre-stimulus alpha phase is interpreted as evidence for fluctuations in state of the visual system prior to stimulus onset affecting the propensity to use prior evidence post-stimulus, at the decision stage. Speculatively, this could occur if prior evidence for or against target presence is periodically transmitted to visual areas, in turn resulting in periodic changes in the baseline from which evidence accumulation begins (Christopher Summerfield & Egner, 2009).

Fluctuations in the influence of expectation on objective decisions were accompanied by fluctuations in subjective confidence. For incongruent reports, subjective violations of expectation were associated with degrees of confidence that tracked the influence of the prior expectation: when perceptual priors exerted greater effects on decision, subjective violations of expectation were associated with greater subjective uncertainty. Moreover, the phase-confidence functions for congruent and incongruent responses were in anti-phase: the phase that predicted greatest uncertainty for incongruent reports also predicted highest confidence for congruent reports. Together, these results extend previous work demonstrating that confidence evolves with the decision variable at early processing stages (Fetsch, Kiani, Newsome, & Shadlen, 2014; Kiani & Shadlen, 2009) by showing that decision and confidence are jointly shaped by top-down influences. As is the case for yes/no decisions, we interpret these results as evidence for biases in the early processing of sensory signals (for example, changes in starting point of evidence accumulation) modulating reported subjective confidence at late stages of the decision-making stream.

Consistent with previous work, we found that alpha phase-modulation of perception is greater with attention than without (Busch & VanRullen, 2010; Landau & Fries, 2012), though here, still present under diverted attention. Critically, while previous evidence has demonstrated alpha-modulation of perceptual hit rate (Busch et al., 2009; Dugué et al., 2011; Landau & Fries, 2012; Mathewson et al., 2009), it has not been possible to ascertain whether changes in hit rate have been driven by changes in sensitivity or bias. Here we implicate alpha oscillations in biasing perceptual decisions, but not increasing sensitivity. Critically, the influence of alpha phase on decision is modulated by expectations. Our data also extend previous research that has revealed that the influence of expectation on decision is predicted by pre-stimulus beta-band power over both motor (de Lange et al., 2013) and somatosensory (van Ede et al., 2010) cortices, as well as by BOLD responses in a range of cortical areas (Hesselmann, Kell, & Kleinschmidt, 2008; Hesselmann, Sadaghiani, Friston, & Kleinschmidt, 2010; Rahnev, Bahdo, de Lange, & Lau, 2012; Christopher Summerfield & Koechlin, 2008). Pre-stimulus signals biasing decision at early stages of visual processing (i.e. in sensory cortices) has not, to our knowledge, been shown before. Our results therefore support an early, and critically, rhythmic, influence of expectations on decision.

Top-down influences are increasingly modeled within Bayesian frameworks (Clark, 2013; Daunizeau et al., 2010; J. Hohwy, 2013; Kersten et al., 2004; Ma, Beck, Latham, & Pouget, 2006; Mathys et al., 2014). Here, perception is described as a Bayesian inference on sensory causes. A core tenet of these frameworks is that the prior probability of sensory causes will constrain inference accordingly, and so probable or ‘expected’ sensory causes are more likely to be chosen and thus perceived (Knill & Pouget, 2004; Lee & Mumford, 2003; Spratling, 2008; Yuille & Kersten, 2006a). A plausible implication of this view is that such prior probabilities should be reflected in the state of the brain in the pre-stimulus period. Consistent with this, we have shown that the influence of priors on decision oscillates with pre-stimulus alpha phase.

One possible explanation for these findings is that alpha oscillations orchestrate the communication of prior expectations to visual cortex. On this view, rhythmic influences of expectation on decision threshold would reflect fluctuations in the prior probability of the reported perceptual decision. However, an alternative view is that our results reflect fluctuations in the weighting of priors on decision, rather than the prior probability itself. On this alternative view, alpha phase reflects the attentional state of the system, consistent with previous theoretical work (Jensen, Bonnefond, & VanRullen, 2012; Palva & Palva, 2007) , so that priors are assigned a greater weight on perceptual decision when sensory signals are expected to be unreliable. Here, perceptual expectations would increase or decrease the excitability of relevant neural populations, or gain, according to whether a target is expected to appear or not. In both cases, pre-stimulus occipital alpha phase modulates the relative weighting of prior expectations and sensory data, however our data cannot discriminate between these two views, and we leave this question open to future research.

In summary, we have described evidence indicating a periodic influence of perceptual priors on both objective (detection) and subjective (confidence) decisions, predicted by the phase of pre-stimulus occipital alpha oscillations. This rapid and periodic alternation between top-down and bottom-up influences in visual areas extends existing data implicating alpha oscillations in top-down processing (von Stein et al., 2000). Together, our data suggest that alpha oscillations may periodically transmit perceptual priors, and in turn reveal a plausible neural mechanism by which prior information may subserve top-down modulation of early visual processing: alpha oscillations may orchestrate the reciprocal exchange of predictions and prediction errors.

6

Functional Network Underlying Top-Down Influences on Confidence

It is clear that prior expectations shape perceptual confidence, yet how this occurs post-stimulus is unknown. Here we recorded fMRI data while participants made perceptual decisions and confidence judgements, controlling for potential confounds of attention. Results show that the relationship between expectations and subjective confidence increases BOLD activity in right inferior frontal gyrus (rIFG). Specifically, rIFG is sensitive to the discrepancy between expectation and decision (mismatch), and, crucially, higher mismatch responses are associated with lower decision confidence. Connectivity analyses revealed the source of top-down influences on confidence to be frontal areas right orbitofrontal cortex (OFC) and bilateral frontal pole (FP), and the source of sensory signals to be occipital pole. Altogether, our results indicate that predictive information is integrated into subjective confidence in rIFG, and reveal an occipital-frontal network that constructs confidence from top-down and bottom-up signals. This interpretation was further supported by exploratory findings that the white matter density of occipital pole and OFC predicted their respective contributions to the construction of confidence. These findings advance our understanding of the neural basis of subjective perceptual processes by revealing a functional network that integrates prior beliefs into the construction of confidence.

6.1 Introduction

Perception is increasingly being seen as an active process, in which current or future sensory states are inferred from predictive information (Bar, 2007; Beck & Kastner, 2009; Engel et al., 2001; Fiser et al., 2010; Gilbert & Li, 2013; Lee, 2002). These predictions can be modelled in Bayesian terms as prior beliefs, which bias perceptual inference towards solutions that are *a priori* more likely in a given context (Bülthoff, Bülthoff, & Sinha, 1998; Seriès & Seitz, 2013; Trapp & Bar, 2015). Predictions, or priors, can have striking effects on perception, especially under high sensory uncertainty. For example, ambiguous rotational motion can be subjectively disambiguated by prior exposure to rotation direction, such that a rotation direction is perceived despite none existing in the physical stimulus (Maloney, Dal Martello, Sahm, & Spillmann, 2005). In laboratory conditions, such behavioural effects of prediction are typically accompanied by increases in BOLD and ERP amplitude, as well as evoked gamma power, over sensory (Kouider et al. 2015; Egner et al. 2010; Saaltink et al. 2015; Kok et al. 2011; Jiang et al. 2013; Wacongne et al. 2011; Bauer et al. 2014) and decision-related (Bubic et al., 2009) brain regions - a 'prediction error' response profile that reflects the discrepancy between internal templates and perceptual content.

The perceptual content that forms the basis of our visual experience is accompanied by a degree of subjective confidence. Confidence reflects the estimated success of a perceptual choice, and can be seen as a gate for post-perceptual processes, such as learning and belief-updating (Nassar, Wilson, Heasly, & Gold, 2010; Yeung & Summerfield, 2012). The communication of decision confidence can also facilitate group decision-making (Bahrami et al., 2010). Yet, while subjective confidence is an integral part of perceptual experience that can be easily probed in human subjects (Fleming & Lau, 2014; Overgaard & Sandberg, 2012; Sandberg et al., 2010; Seth et al., 2008; Wierzchoń et al., 2014), the construction of confidence remains poorly understood.

It is clear that confidence increases with evidence in support of the decision (Fetsch, Kiani, & Shadlen, 2015; Gherman & Philiastides, 2015; Hebart et al., 2014; Yeung & Summerfield, 2012). Decision and confidence are thought to evolve together until the first-order, objective decision has been made (Kepecs & Mainen, 2012; Ratcliff & Starns, 2009), and accordingly, there exists strong evidence for a common sensory signal underlying both types of report (Fetsch et al., 2014; Kiani et al., 2014; Kiani & Shadlen, 2009). Surprisingly, there has been much less research that considers the role of prior expectations on subjective confidence. There is converging behavioural evidence for subjective confidence increasing with prior evidence in favour of the associated choice (see Chapters 4 and 5, and Aitchison, Bang, Bahrami, & Latham, 2015; Meyniel, Schlunegger, & Dehaene, 2015), but the neural substrates of this have remained unexplored.

Here we aimed to identify brain regions in which prior perceptual expectations are integrated into confidence judgements. Based on previous work, we reasoned that confidence should be high when decisions are supported by prior knowledge, that is, when the discrepancy between expectation and perceptual decision is low. We therefore sought to identify brain regions that, first, are sensitive to both prediction error and confidence, and second, in which confidence is negatively associated with prediction error. In such a region, confidence would be associated with the mismatch between internal templates and perceptual report.

We further hypothesised that regions found to integrate prior expectations into confidence judgements (as described above) should be functionally connected with two information sources: one that represents the decision evidence, or sensory information; and one that represents the prior expectation. As confidence increasingly depends on prior expectations, functional connectivity with the source of the priors and sensory signals relevant to these judgements should increase.

6.2 Materials and Methods

6.2.1 Participants

The study was approved by the Brighton and Sussex Medical School Research Governance and Ethics Committee. Twenty-four healthy, English speaking and right-handed subjects were tested (age 19-34, mean age = 25 years, 13 females). Data from five participants were excluded: two whose thresholding failed (see section ‘Staircases’, Gabor hit rate = 2%, visual search d’ = -0.1); one for revealing abnormal vision only after scanning (and whose estimated contrast thresholds were accordingly > 2SD from the mean); one for excessive head movement in the scanner such that their T1 scan was unusable; and one for failing to respond on 33% of trials (relative to a mean of 3%). This left 19 participants with normal or corrected-to-normal vision for analysis. All participants gave informed, written consent and were reimbursed £50 for their time.

6.2.2 Procedure

The experiment was conducted over three sessions at least 2 hours apart (no participant completed all three on a single day). In session one informed consent was obtained. Participants were trained on all tasks before scanning, which consisted of on-screen instructions, followed by a minimum of 10 practice trials of each task. Participants were encouraged to continue training until the task was well understood and response mappings learned.

To equate performance accuracy across conditions and subjects, participants subsequently completed three staircase procedures in the scanner but without acquiring echoplanar images (EPIs). Next, two 17 minute runs of experimental trials were completed while EPI scans were acquired. Session two did not include a training component but was otherwise identical to session one. Session three consisted of: 10 minutes for T1 acquisition; 15 minutes of retinotopy (data from which is not used in this paper); and, time permitting, one more experimental run. After three sessions participants were compensated for their time and debriefed.

6.2.3 EXPERIMENTAL DESIGN

The paradigm used in the present study is adapted from that used in Chapter 4. The visual display was identical in all sections of the experiment (training, staircase and experimental). It consisted of a central visual search array and the presence or absence of a to-be-detected, Gabor patch in the periphery (see figure 6.1 and subsection 'Trial Sequence').

In experimental trials, the principal task was Gabor detection and two factors were orthogonally manipulated: prior expectations of Gabor presence and attention to Gabor detection. Expectations were manipulated block-wise, by changing the probability of target Gabor presentation (P(Gabor present) = .25,

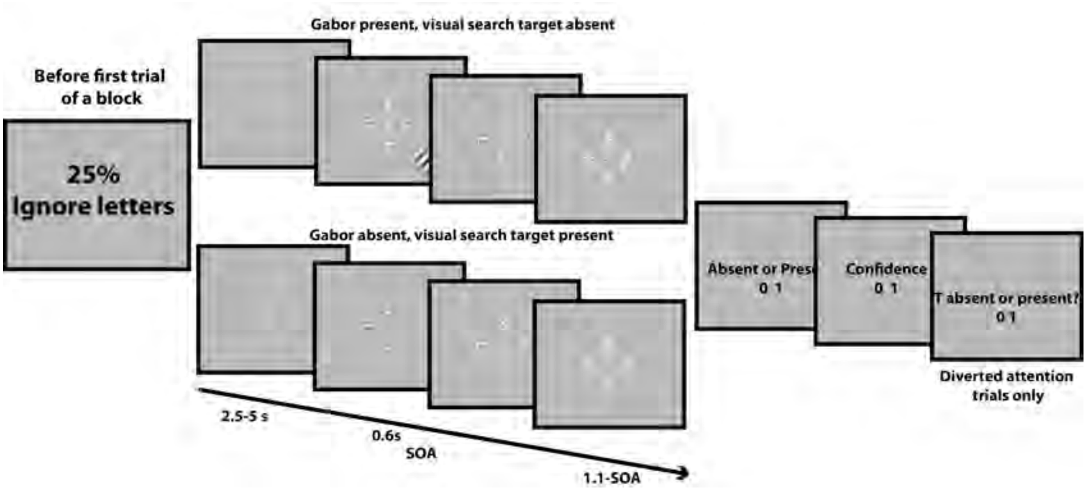

Figure 6.1. Trial sequence.

Blocks began with instructions signalling the expectation and attention condition. Here, the block induces an expectation of target absence (Gabor presentation is unlikely) and the central visual search task should be ignored (diverted attention condition). On each trial a visual search target T was either absent (top) or present (bottom) with 50% probability. On each trial a target Gabor was either present (top) or absent (bottom) with probability determined according to condition. Response cues followed the offset of the stimuli. Staircase trials were identical, except there was no condition-specific instruction at the beginning and only task-relevant response cues were presented.

.50 or .75). The P(Gabor present) = .25 condition induced an expectation of Gabor absence, whereas the P(Gabor present) = .75 condition induced an expectation of Gabor presence. The P(Gabor present) = .50 condition acted as a control (flat prior). Attention was manipulated by instructing participants to either perform or ignore a visual search task presented concurrently to the Gabor target. This task consisted of detecting target ‘T’s amongst an array of distracter ‘L’s. Performing both tasks concurrently diverted attention from the Gabor detection task, allowing us to separate effects of expectation from those of attention.

These conditions were manipulated block-wise, in groups of 12 trials. Each condition occurred once per scanning run in fully counter-balanced order. Before each experimental block began participants were informed of both the expectation and attention condition via the presentation of an instruction screen presented for 10 seconds. Participants were instructed to always maintain fixation at a central cross.

6.2.4 Trial sequence

The trial sequence was identical for training, staircasing and experimental trials and is shown in figure 6.1. Only instructions varied (see ‘Experimental design’). Trials began with a white fixation cross of random duration between 2.5 and 5 seconds. Next, a visual search array appeared, which consisted of seven letters: all white, capital ‘L’s (50% chance), or a white, capital ‘T’ replacing an ‘L’ (50% chance). All letters were equidistant from fixation and took an independently random orientation. These were subsequently masked by a matching array of ‘F’s to increase task difficulty. In total the visual search array was present for 1.1 seconds. The stimulus onset asynchrony (SOA) between target and masking arrays was titrated for each participant such that accuracy was at 78% (see Staircases).

On some trials a near-threshold (see section Staircases) peripheral Gabor patch (orientation = 135°, phase 45° on 50% of trials, 225° on 50% of trials, sf = 2c/°, Gaussian SD = 30) was additionally presented. On these trials the stimulus appeared at the same time as the visual search array. To minimise attentional capture it was presented over 0.6 seconds in a Gaussian time envelope so that it had a gradual onset and offset. Stimulus contrast was titrated to equate performance across levels of attention and participants at 78% accuracy (see Staircases).

The interval between offset of the masking array and onset of response prompts was jittered during scanning only (i.e. experimental trials) to minimise motor cortex activity reflecting response anticipation. Jitter was randomly selected from the discrete values 1.3s:0.3s:3.1s.

Response prompts were presented at the end of the trial. The first prompt referred to the Gabor detection task. ‘Absent’ responses were recorded by pressing the outer left key and ‘present’ responses, the outer right key. This prompt was presented on all trials except those of the visual search staircase procedure (only visual search performed). The second prompt asked whether participants guessed (inner left) or were confident (inner right) in their Gabor detection response (not presented on staircasing trials). The third prompt was only presented on trials where participants performed the visual search task. This asked whether the visual search target ‘T’ was absent (outer left) or present (outer right). Response prompts remained onscreen for 2 seconds and responses were coded as missed trials if no response was given within the allowed time.

6.2.5 Staircases

Prior to each experimental session, three separate adaptive 1-up-3-down psychophysical staircase procedures (9 reversals) were completed in the scanner. Trials were identical to those in staircase trials (see Trial structure) except: there was no manipulation of attention or expectation; the Gabor was always present, but randomly oriented either 45° to the left or to the right; the Gabor task was 2AFC orientation discrimination instead of target detection; confidence ratings were not requested.

Staircase 1 titrated Gabor contrast to achieve 78% accuracy under full attention. Initial contrast was 1.5%. The visual search array was masked after 0.5 seconds. Participants were instructed to ignore the visual search array but still fixate centrally.

Staircase 2 titrated the SOA between the visual search array and masking array to set performance at 78% (in the visual search task). Initial SOA was 500ms. Participants ignored the 2AFC task and performed the visual search task. Here, the ignored Gabor was presented at the contrast acquired in staircase 1.

Staircase 3 titrated Gabor contrast to achieve 78% accuracy (in Gabor detection) under diverted attention. Initial contrast was set at that obtained in staircase 1 and visual search SOA was set at the value obtained by staircase 2. Here, participants performed both the Gabor and the visual search tasks. The visual search SOA was set at the value obtained in the previous staircase and initial contrast was set at that obtained in the first and titrated over the course of the staircase to obtain the diverted attention contrast level.

6.2.6 Statistical analyses

Gabor detection sensitivity and decision threshold were quantified by computing type 1 signal detection theoretic (SDT) measures *d'* and *c* respectively (see Chapter 3 for more detail). These are computed by classifying trials as hits (*h*), misses (*m*), false alarms (*fa*) or correct rejections (*cr*). Then,

$$HR = \frac{\sum h}{\sum h + \sum m} \; and \; FAR = \frac{\sum fa}{\sum fa + \sum cr}$$

so that

$$d' = \; Z(HR) - Z(FAR) \; and \; c = -\frac{Z(HR) + Z(FAR)}{2}$$

where Z is the inverse cdf of the normal distribution.

Confidence was computed by calculating the proportion of trials on which each subject reported ‘confident’.

We also estimated metacognitive bias. In signal detection theory, we can represent confidence thresholds that fall on either side of the type 1 criterion, representing the additional evidence needed to report “confident no” or “confident yes”. These thresholds reflect metacognitive bias. These thresholds, are defined as follows:

$$C'_1 + = \left|\frac{c+}{d'+} - \frac{c}{d'}\right| \; and \; C'_1 - = \left|\frac{c-}{d'-} - \frac{c}{d'}\right|$$

The plus and minus signs correspond to metacognitive bias for “yes” and “no” responses respectively. The values *c*+ and *d’*+ are computed by reclassifying hits, misses, false alarms and correct rejections according to reported confidence. Confident “yes” reports are reclassified as simply “yes” and all others as “no”. From these, we obtain *c*+ and *d+’* from the standard equations for *d’* and *c*. This process is repeated for “no” responses by setting “no” to be “confident no” and “yes” otherwise. These values of metacognitive bias should be independent of decision accuracy. High values correspond to thresholds that extend far from the type 1 criterion, meaning that metacognitive bias is conservative (more likely to report “guess”). The reverse applied for small values.

In the present Chapter, two participants had one extreme value of metacognitive bias (> 20 times larger than their other estimates), resulting from dividing by a small response-conditional *d’*. For these participants, the extreme value was replaced with the subject’s mean across the other 11 estimates.

Behavioural and follow-up statistical tests were run on JASP (Love, et al., 2015). When the null hypothesis was predicted, Bayesian t-tests and repeated-measures ANOVAs implemented the JASP default Cauchy prior of 0.7 HWHM centered on zero. All results presented were robust to reasonable adjustments of this value. Bayes factors greater than 1/3/10/100 are respectively interpreted as showing insensitive/moderate/strong/very strong evidence for the alternative hypothesis (Kass & Raftery, 1995). Bayes factors less than the reciprocal of these values are given the same labels, but refer to the null hypothesis.

Unless otherwise stated, all repeated-measures ANOVA results met the assumption of sphericity. Where sphericity was violated, corrected degrees of freedom and *p*-values are presented. The Greenhouse-Geisser correction is used for small violations (ε < .75) and the Huynh-Feldt correction for large violations (ε > .75).

6.2.7 MRI ACQUISITION AND PRE-PROCESSING

Functional T2* sensitive echoplanar images (EPIs) were acquired on a Siemens Avanto 1.5T scanner. Axial slices were tilted to minimise signal dropout from frontal and occipital cortices. 34 2mm slices with 1mm gaps were acquired (TR = 2863ms, TE = 50ms, FOV = 192mm x 192mm, Matrix = 64 x 64, Flip angle = 90°). Full brain T1-weighted structural scans were acquired on the same scanner and were composed of 176 1mm thick sagittal slices (TR = 2730ms, TE = 3.57ms, FOV = 224mm x 256mm, Matrix = 224 x 256, Flip angle = 7°) using the MPRAGE protocol.

Functional runs, each lasting 17 minutes, were collected per scanning session. Images were processed using SPM8 software (http://www.fil.ion.ucl.ac.uk/spm/software/spm8/). The first four functional volumes of each run were treated as dummy scans and discarded. Images were pre-processed using standard procedures: anatomical and functional images were reoriented to the anterior commissure; images were slice-time corrected with the middle slice used as the reference; EPIs were aligned to each other and co-registered to the structural scan by minimising normalised mutual information. Next, EPIs were spatially normalised to MNI space using parameters obtained from the segmentation of T1 images into grey and white matter. Finally, spatially normalised images were smoothed with a Gaussian smoothing kernel of 8mm FWHM.

6.2.8 FMRI STATISTICAL ANALYSIS

At the participant level BOLD responses were time-locked to the onset of the visual search array (which appeared at the same time as the Gabor, if present), enabling us to examine BOLD responses to both target present and target absent trials. BOLD responses were modelled in a GLM with regressors and their corresponding temporal derivatives for each combination of the following factors: Attention (full, diverted), Expectation (25%, 50%, and 75%), Stimulus (target present, target absent), Report (yes, no) and Confidence (confident, guess). If a certain combination of factors had no associated trials for a particular participant, that regressor was removed from the participant's first level model and contrast weights rescaled.

The reliability of the regression weights was maximised by entering data from all runs and sessions together, increasing the trial count per regressor. To avoid smearing artefacts, no band-pass filter was applied. Instead, low-frequency drifts were regressed out by entering white matter drift (averaged over the brain) as a nuisance regressor (Law et al., 2005). Nuisance regressors representing the experimental run and six head motion parameters were also included.

Comparisons of interest were tested by running one-sample *t*-tests against zero at the participant level, then running group-level paired *t*-tests on the one-sample maps. Unless otherwise stated, all contrasts at the group level were run with peak thresholds of $p < .001$ (uncorrected) and corrected for multiple comparisons at the cluster level using the FDR method.

We wanted to control for possible confounds between reaction speed and confidence (which correlate, see e.g. Grinband, Hirsch, & Ferrera, 2006; Petrusic & Baranski, 2003), and between individual or condition-wise differences in Gabor contrast and confidence (which correlate, Rahnev et al., 2011). To do this, a control GLM was computed. Here, each regressor was parametrically modulated by both Gabor contrast and reaction time. By design, in this model confidence was independent of reaction time and BOLD amplitude was independent of individual and condition-wise differences in stimulus contrast. The Results section reports analyses on our main model, i.e. the model without regressors for Gabor contrast and reaction speed. We did this because the control model has a four-fold increase in number of regressors, reducing statistical power. Nonetheless, all GLM analyses were replicated under our control model when using a peak threshold of $p < .005$. Crucially, all results under rIFG were also replicated when using a peak threshold of $p < .001$.

Functional ROIs were defined using the MarsBaR 0.42 toolbox (http://marsbar.sourceforge.net/download.html). Anatomical areas showing significant differences in BOLD were identified using the SPM Anatomy toolbox (Eickhoff et al., 2005) and Brodmann areas were identified using MRIcro (Rorden & Brett, 2000). Results of whole-brain analyses were plotted onto glass brains using MATcro (now called MRIcroS, https://www.nitrc.org/plugins/mwiki/index.php/mricros:MainPage).

6.2.9 Psychophysiological interaction (PPI) analysis

The psychophysiological interaction (PPI) analysis was performed using the CONN functional connectivity toolbox (http://web.mit.edu/swg/software.htm). The GLM comprised regressors for attention condition (full/diverted), confidence (confident/guess) and expectation-response congruence (congruent/neutral/incongruent). Nuisance regressors were identical to those used in the GLM on BOLD. Again, the signal was not band-pass filtered but instead the mean WM drift was entered as a nuisance regressor. The data were denoised by regressing out signal from white matter, from CSF and from each individual condition, plus signal associated with all nuisance regressors. The PPI was run on univariate regression weights to identify effective connectivity between a functionally defined seed (rIFG) and remaining voxels. These weights were examined in a second level model which used an uncorrected peak threshold of $p < .005$ and FDR cluster corrected threshold of $p < .05$.

6.2.10 Voxel-based morphometry (VBM)

T1-weighted structural scans were reoriented to the anterior commissure and segmented into grey matter (GM), white matter (WM) and CSF. These were normalised to MNI space using DARTEL with SPM defaults and a Gaussian smoothing kernel of 8mm FWHM (Ashburner & Friston, 2000). White matter and grey matter images were separately compared across participants in a multiple regression with age and total intracranial volume (GM + WM + CSF) as nuisance regressors. Gender was not included because this resulted in multicollinearity between regressors (older participants were more likely to be male). Unless reported otherwise, clusters reported as significantly correlating with behaviour survived voxel-wise FWE correction.

6.4 RESULTS

6.3.1 EXPECTATIONS LIBERALISE DECISIONS AND ATTENTION INCREASES CONTRAST SENSITIVITY

Our first analyses confirmed the efficacy of our paradigm. To equate difficulty across attention conditions and participants, adaptive psychophysical staircases identified the stimulus contrast required for 78% accuracy on the Gabor detection task (see Methods subsection Staircases). Comparing the acquired contrasts in the full and diverted attention conditions revealed that contrast thresholds were significantly lower under full than diverted attention, $t(19) = 2.95$, $p = .014$, 95%CI [0.50%, 2.31%], $d_z = 0.70$ (fig. 6.2A). Thus, our paradigm successfully manipulated attention.

To ensure that our staircase procedure successfully equated detection sensitivity *d'* across conditions we ran a within-subjects Attention (full, diverted) x Expectation (25%, 50%, 75%) ANOVA. This revealed no significant difference between *d'* under full (M = 1.06, SE = 0.14) and diverted (M = 1.21, SE = 0.20) attention conditions, $F(1,18) = 0.34$, $p = .569$, $\eta_p^2 = .02$ (fig. 6.2B), and was corroborated by a Bayesian repeated-measures ANOVA of the same design that revealed moderate evidence for the null hypothesis (BF = 0.240). There was also no significant effect of Expectation on *d'*, $F(2,36) = 0.70$, $p = .505$, $\eta_p^2 = .04$, BF = 0.07 (strong evidence for the null) and no significant interaction term $F(2,36) = 0.76$, $p = .476$, $\eta_p^2 = .04$, BF = 0.016 (strong evidence for the null). Our staircases therefore successfully equated *d'*.

To determine whether we had successfully manipulated priors, we compared signal detection theoretic decision thresholds (*c,* see Methods) across

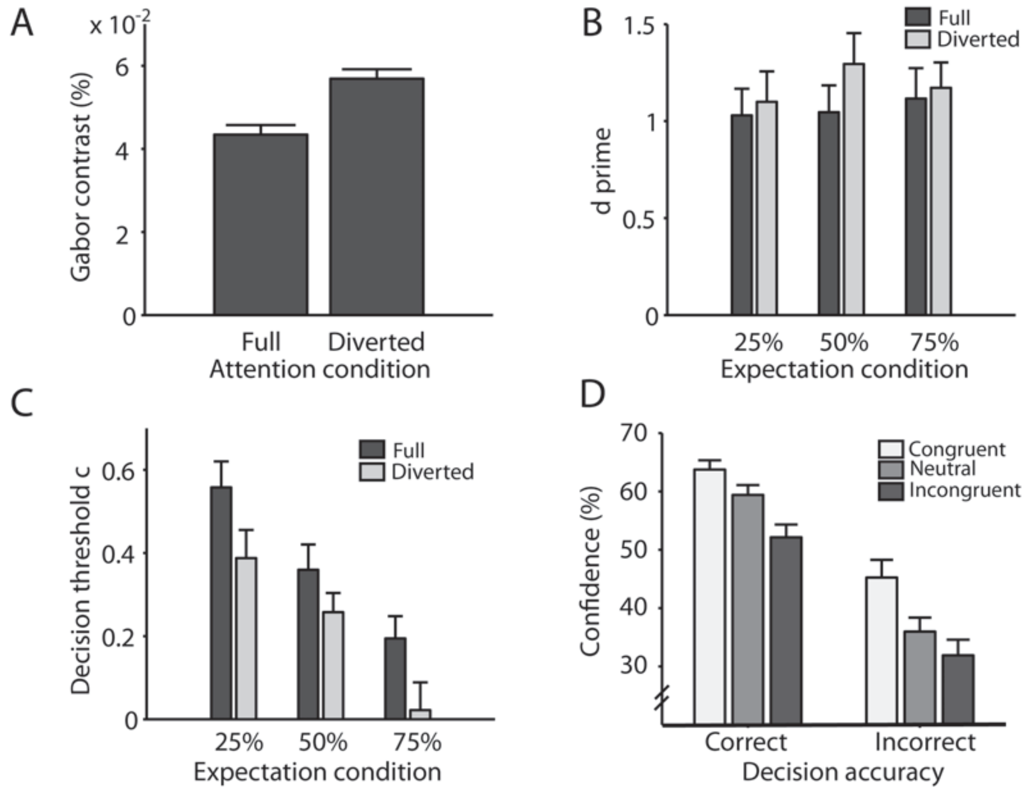

Figure 6.2. Behavioural effects of expectation and attention on objective and subjective decision-making.

A. Stimulus contrast as a function of attention condition. To achieve 78% correct on the Gabor detection task contrast had to be higher under diverted than full attention.
B. Detection sensitivity *d'* as a function of expectation and attention condition. No significant differences were found.
C. Decision threshold c as a function of expectation and attention condition. Independently of attention, bias towards reporting ‘yes’ (lower values of *c*) increases with the prior probability of Gabor presence.
D. Confidence as a function of expectation-report congruence and decision accuracy. Independently of accuracy, confidence is higher for congruent that incongruent reports. Untransformed data are presented for illustrative purposes.
Error bars represent within-subjects SEM.

expectation conditions (see Chapter 4, and also de Lange et al., 2013; Morales et al., 2015). As the expectation of Gabor presence over absence increases, decision threshold should become increasingly biased towards ‘yes’ responses (i.e. liberalised, shown by smaller values of *c*). This was confirmed in a within-subjects Attention (full, diverted) x Expectation (25%, 50%, 75%) ANOVA, $F(1.65, 29.72) = 18.10$, $p < .001$, $\eta_p^2 = .50$. LSD post-hoc tests revealed a greater bias towards reporting ‘yes’ in the 50% (neutral) than the 25% (expect absent) condition, $p = .010$, $d_z = 1.15$, and greater still in the 75% (expect present) than the 50% (neutral) condition, $p < .001$, $d_z = 1.39$ (fig. 6.2C). We found no evidence for attentional effects on decision threshold, $F(1, 18) = 3.38$, $p = .083$, $\eta_p^2 = .16$, and no Expectation x Attention interaction, $F(2, 36) = 0.37$, $p = .693$, $\eta_p^2 = .020$. Summarising these results, our design successfully independently manipulated attention and expectation, while keeping detection sensitivity constant across conditions.

6.3.2 Expectations increase confidence

In Chapters 4 and 5 it was shown that subjective confidence increases when perceptual decisions are congruent with prior expectations. On this basis, we hypothesised that confidence would relate to prediction error signals. To determine whether we had replicated this behavioural result, we compared confidence for perceptual decisions that were congruent with expectations against those that were incongruent. Congruent responses are 'yes' reports in the 75% (expect present) condition and 'no reports in the 25% (expect absent) condition. The reverse applies for incongruent responses.

The percentage of high confidence trials were Z-transformed (because otherwise confidence is bounded by 0 and 1) and subjected to an Attention (full, diverted) x Accuracy (correct, incorrect) x Congruence (congruent, neutral, incongruent) repeated-measures ANOVA. Participants appropriately showed lower confidence for incorrect than correct reports, $F(1,18) = 5.70$, $p = .028$, $\eta^2 = .241$. Confidence was also higher for attended than unattended targets $F(1,18) = 5.27$, $p = .034$, $\eta^2 = .226$. Importantly, confidence increased with expectation-response congruence, $F(1.10,19.86) = 6.67$, $p = .016$, $\eta^2 = .270$ (untransformed data plotted in figure 6.2D). Congruence additionally interacted with attention, $F(2,36) = 6.51$, $p = .004$, $\eta^2 = .266$, such that inattention only lowered confidence when participants had an informative prior (congruent reports $p = .023$, incongruent reports, $p = .006$ vs. neutral reports, $p = .280$). There were no other significant main or interaction effects (all $p > .107$, all $\eta^2 < .138$).

Next we wanted to determine whether this effect reflected changes in metacognitive bias. For each level of attention and expectation, we computed response-specific confidence thresholds over the type 1 SDT model (see Methods and Section 3.6.1). We wanted to determine whether confidence thresholds for congruent decisions were closer to the type 1 criterion than incongruent decisions. Values that are closer to zero indicate that the participant reports “confident” with less evidence, thereby indicating a more liberal metacognitive bias.

As shown in figure 6.3, we found a significant expectation by report interaction, $F(2,36) = 17.16$, $p < .001$, $\eta^2 = .488$, such that when participants reported “yes”, metacognitive bias linearly liberalised with increasing prior probability of target presence, $F(1,18) = 6.24$, $p = .022$, $\eta^2 = .257$, and when they reported “no”, metacognitive bias linearly liberalised with prior probability of absence, $F(1,18)$

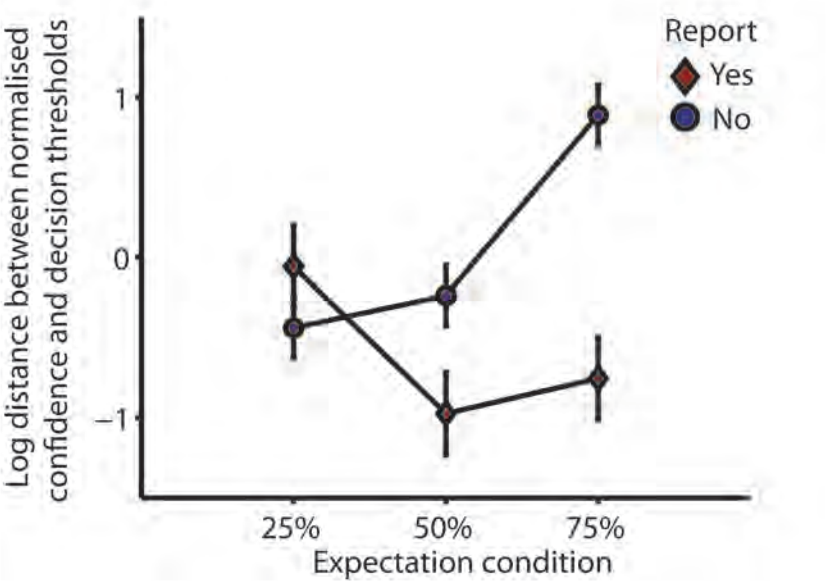

Figure 6.3 Response-specific metacognitive bias as a function of expectation.

The red diamonds depict the distance of the normalised confidence threshold for ‘yes’ responses from the normalised type 1 threshold. This threshold is squeezed towards criterion with increasing probability of target presence, representing more liberal confidence. Similarly, for targets reported as absent (blue circles), the normalised confidence threshold is closer to the normalised criterion with increasing probability of target absence. Therefore, expectation-congruent responses require less type 1 evidence than incongruent responses to be reported with confidence. Error bars represent within-subject SEM.

= 24.00, p < .001, η^2 = .571. These results indicate that over and above effects on type 1 decision threshold, expected percepts may require less type 1 evidence to be reported with high confidence.

Finally we asked whether the congruency effect on metacognition found in Chapter 4 was replicated. We ran the same analysis as that above, but on type 2 *D'* and *meta-d'/d'*. Results showed that while type 2 *D*' increased with expectation-response congruence independently of attention, $F(2,36) = 14.25$, $p < .001$, η^2 = .952, no effect on *meta-d'/d'* was found (Expectation x Report, p = .940, Expectation x Report x Attention, p = .284). Thus, we partly replicated findings from Chapter 4.

6.3.3 Two forms of congruency

To unravel the neural correlates of predictive influences on confidence, we first needed to identify brain regions sensitive to perceptual expectations. We predicted, based on previous work, that areas sensitive to perceptual expectations would exhibit an increased BOLD amplitude for trials on which expectations were violated (Egner et al., 2010; Jiang et al., 2013; Kok et al., 2011; Kouider et al., 2015; St. John-Saaltink et al., 2015).

There are two possible ways to define expectancy violations here. Because the experimental design used near-threshold stimuli, leading to potential dissociations between percept and physical stimulus presentation, violations could occur with respect to either physical stimulus presentation, or perceptual report. We term the neural correlates of these types of incongruence $PE_{STIMULUS}$ and PE_{REPORT} respectively. The former reflects the BOLD response to discrepancy between internal templates and stimulus presentation, whereas the latter reflects the BOLD response to discrepancy between internal templates and participants' reported percept. $PE_{STIMULUS}$ is most often observed at lower levels of the perceptual hierarchy (Chennu et al., 2013; Jiang et al., 2013; Kok et al., 2011), whereas the decision-related PE_{REPORT} signals are often reported in higher-level, decision-related areas (Bubic et al., 2009), though they can be observed in visual cortex as well (Pajani, Kok, Kouider, & de Lange, 2015).

6.3.4 Representation of PE$_{\text{STIMULUS}}$ in visual cortex

In our first analysis, we searched for regions that are sensitive to discrepancies between expectation and stimulus presentation (PE$_{\text{STIMULUS}}$) over whole brain. To do this, we computed the contrast unexpected stimulus presentation > expected stimulus presentation. Target presence is expected in the 75% condition but unexpected in the 25% condition. Target absence is expected in the 25% condition but unexpected in the 75% condition. Our analysis identified one PE$_{\text{STIMULUS}}$-sensitive area in contralateral occipital cortex (V1 to V3, BA18, peak MNI x = -12, y = -80, z = 22, Z$_{\text{peak}}$ = 4.09, 0.66cm^3, cluster p_{FDR} = .350, p_{uncorr} = .023) and one on the ipsilateral side (V1 to V3, BA18, peak MNI x = 8, y = -80, z = 18, Z$_{\text{peak}}$ = 3.99, 1.01cm^3, cluster p_{FDR} = .205, p_{uncorr} = .007). Neither of these clusters survived cluster-level correction, so they will not be considered beyond this point. They are presented to simply to show consistency with previous studies, in which statistical power was improved by constraining the analysis with functional localisers (Jiang et al., 2013; Kok et al., 2012, 2011; Larsson & Smith, 2012; Smith & Muckli, 2010) .

The whole-brain contrast PE$_{\text{STIMULUS}}$, attended > PE$_{\text{STIMULUS}}$, unattended yielded no significant or marginally significant clusters, indicating no evidence for a PE$_{\text{STIMULUS}}$ × attention interaction.

Using a peak threshold of p < .005 both of these analyses were replicated under our control model, which included reaction speed and Gabor contrast as parametric modulators (unexpected > expected, contralateral: p_{FDR} = .446, p_{uncorr} = .014, ipsilateral: p_{FDR} = .446, p_{uncorr} = .011).

6.3.5 Regions representing PE$_{\text{REPORT}}$

Next, we searched for regions whose BOLD response reflects the discrepancy between expectation and perceptual report (PE$_{\text{REPORT}}$). Expectation-congruent reports are 'yes' responses in the 75% (expect present) condition and 'no' responses in the 25% (expect absent) condition. The reverse applies for expectation-incongruent reports. These definitions differ from those in the previous analysis, because they consider perceptual report instead of stimulus presence or absence.

The contrast expectation-incongruent report > expectation-congruent report was computed over whole-brain. This revealed eight significant clusters, distributed throughout cortex (figure 6.4A and table 6.1). Our control analysis revealed that this difference was not driven by differences in Gabor contrast or reaction speed (all remained significant after cluster correction at the $p < .05$ level).

We found no significant clusters for the reverse contrast, even with a more liberal peak threshold of $p < .005$ uncorrected.

Regions exhibiting a $\mathrm{PE}_{\mathrm{REPORT}}$ pattern should show heightened BOLD for incongruent responses irrespective of whether that response was a ‘yes’ or a ‘no’ (Kok et al., 2011). To test this in the above ROIs, median regression coefficients were extracted as a function of attention, expectation and report, and subjected to separate repeated-measures ANOVAs. Results are depicted in figure 6.4B and statistics are presented in table 6.2. All regions exhibited a significant $\mathrm{PE}_{\mathrm{REPORT}}$ response for both ‘yes’ and ‘no’ judgements, except middle orbital gyrus and left inferior frontal gyrus. As a result these are not considered regions representing $\mathrm{PE}_{\mathrm{REPORT}}$. All significant results here were replicated (at least at marginal significance) under our control model (for rIFG, our critical region, $p_{\mathrm{FDR}} = .044$). Results were fully replicated when using a peak threshold of $p < .005$. We have therefore identified six regions signalling $\mathrm{PE}_{\mathrm{REPORT}}$. These are: right middle temporal gyrus (rMTG); right superior medial gyrus (rSMG), right inferior frontal gyrus (rIFG); right angular gyrus (rAG); and bilateral supramarginal gyrus (SG). These results implicate this set of regions as having sensitivity to the discrepancy between perceptual expectations and perceptual choice.

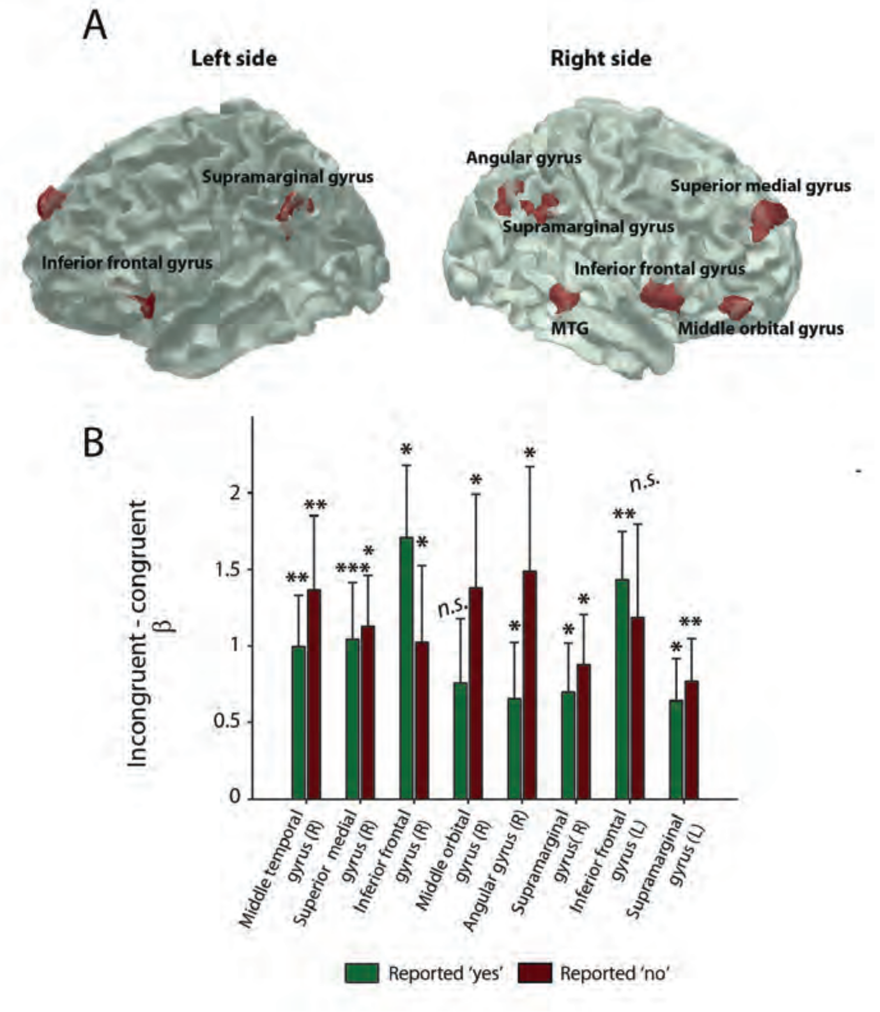

Figure 6.4. Report prediction error.

A. Results of contrast incongruent response > congruent response over whole brain. Only clusters surviving FDR cluster-correction are shown. **B.** PE_{REPORT} (incongruent – congruent), by region and perceptual report. BOLD has been averaged over levels of attention. Stars represent whether PE_{REPORT} is significantly different from zero. Error bars represent SEM * p < .05, ** p < .01, *** p < .001

Table 6.1. Results of whole-brain analysis expectation-incongruent report > expectation-congruent report

Region	*BA*	*Side*	*Volume (cm^3)*	*Peak Z*	p_{FDR}	*Peak MNI*		
						x	*y*	*z*
MTG	*21*	*R*	*2.29*	*4.78*	*.007*	*54*	*-30*	*-2*
SMG	*9/10*	*R*	*4.15*	*4.54*	*< .001*	*12*	*58*	*32*
IFG	*47/48*	*R*	*2.70*	*4.45*	*.004*	*56*	*12*	*-2*
MOG	*47/46*	*R*	*2.08*	*4.33*	*.009*	*40*	*50*	*-6*
AG	*39*	*R*	*1.21*	*3.95*	*.044*	*46*	*-64*	*36*
SG	*40*	*R*	*1.21*	*3.91*	*.044*	*58*	*-40*	*40*
IFG	*47*	*L*	*1.90*	*3.79*	*.012*	*-38*	*26*	*-4*
SG	*40/48*	*L*	*1.60*	*3.75*	*.021*	*-54*	*-46*	*34*

MTG = middle temporal gyrus, SMG = superior medial gyrus, IFG = inferior frontal gyrus, MOG = middle orbital gyrus, AG = angular gyrus, SG = supramarginal gyrus

Table 6.2. Effect of expectation, separately for 'yes' and 'no' reports. Both effects should be significant for the region to be deemed a PE_{REPORT} region

Region	*Reported 'no'*			*Reported 'yes'*			PE_{REPORT}
	F	*P*	η^2	*F*	*p*	η^2	
Middle temporal gyrus	*8.82*	*.008*	*3.29*	*5.83*	*.006*	*.245*	*Yes*
Superior medial gyrus	*8.10*	*.001*	*.310*	*4.46*	*.014*	*.213*	*Yes*
Inferior frontal gyrus (R)	*4.70*	*.015*	*.207*	*3.45*	*.041*	*.162*	*Yes*
Middle orbital gyrus	*1.95*	*.157*	*.098*	*3.42*	*.044*	*.160*	*No*
Angular gyrus	*3.52*	*.040*	*.164*	*4.07*	*.025*	*.185*	*Yes*
Supramarginal gyrus (R)	*4.71*	*.044*	*.207*	*7.17*	*.015*	*.285*	*Yes*
Inferior frontal gyrus (L)	*5.62*	*.008*	*.238*	*2.87*	*.070*	*.137*	*No*
Supramarginal gyrus (L)	*5.39*	*.032*	*.230*	*6.04*	*.005*	*.251*	*Yes*

6.3.6 High confidence is associated with an attenuated $\mathrm{PE}_{\mathrm{REPORT}}$ response in right IFG

Our main hypothesis was that high confidence would be associated with low $\mathrm{PE}_{\mathrm{REPORT}}$. However, confidence can be also influenced by attention (Rahnev et al., 2011) and tracks accuracy (Dienes, 2008; Pleskac & Busemeyer, 2010). To

test whether any $\mathrm{PE}_{\mathrm{REPORT}}$ region represented confidence after controlling for these potential confounds, median regression weights from each $\mathrm{PE}_{\mathrm{REPORT}}$ region were extracted as a function of confidence, attention and decision accuracy. These regression coefficients were then subjected to separate Bayesian repeated-measures ANOVAs. We were looking for regions whose BOLD response (in these regions, representing $\mathrm{PE}_{\mathrm{REPORT}}$) differs with confidence. Note that we could not test for a $\mathrm{PE}_{\mathrm{REPORT}}$ x Confidence interaction because the participant has signalled low confidence yes/no decisions as unreliable, that is, their perception of Gabor presence or absence does not necessarily correspond to their report.

Only one region exhibited a BOLD response (i.e. $\mathrm{PE}_{\mathrm{REPORT}}$ amplitude) that differed as a function of subjective confidence: rIFG. Here, supporting our hypothesis, BOLD amplitude was higher for guess than confident reports (figure 6.5A). Crucially, the analysis revealed substantially more evidence for modelling rIFG BOLD as a function of confidence alone (BF = 13.620) than as a function of just accuracy (BF = 0.877), just attention (BF = 0.711), or as a combination of confidence and any other factors (BF = 0.003 - 2.069, see table 6.3 for summary of results from all ROIs). A frequentist ANOVA gave the same result: a significantly higher BOLD amplitude for guess than confident responses, $F(1,18) = 6.04$, $p = .024$, $\eta^2 = .251$, 95% CI [0.10, 1.28]. These results are depicted in figure 6.5B.

Next, we wanted to confirm that the effect of confidence on rIFG BOLD indeed reflects changes in $\mathrm{PE}_{\mathrm{REPORT}}$. To do this, we restricted our analysis to confident responses and asked whether $\mathrm{PE}_{\mathrm{REPORT}}$ decreases as expectations exert stronger influences on behavioural confidence. This would show that high

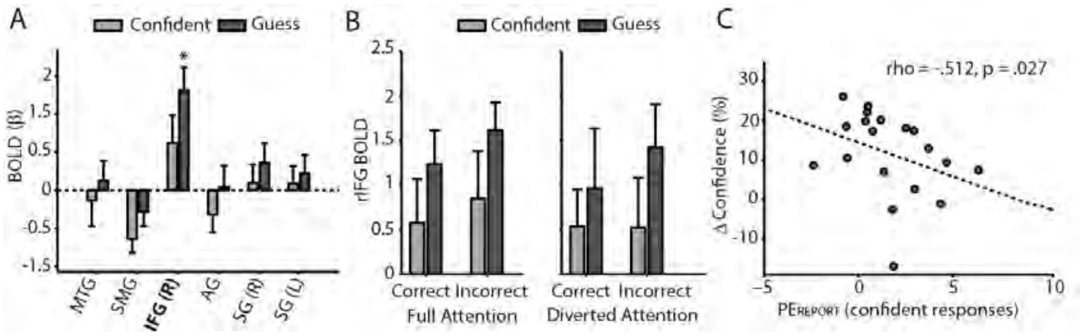

Figure 6.5.The relationship between confidence and report prediction error.

(A) BOLD as a function of confidence in each PE$_{\text{REPORT}}$ region. BOLD is significantly higher for guess than confident responses in rIFG only. **(B)** rIFG BOLD is higher for guess than confidence responses independently of attention and decision accuracy. **(C)** Brain-behaviour correlation. The higher the PE$_{\text{REPORT}}$ response (confident reports only), the less expectations increased confidence. Error bars represent SEM. * *p* < .05, ** *p* < .01, *** *p* < .001

confidence is associated with low PE$_{\text{REPORT}}$ amplitude (i.e. a low expectation-report mismatch response). Furthermore, it would show that our behavioural effect of expectation on confidence is reflected in rIFG BOLD.

To test this, we calculated the percentage increase in confidence when reports were congruent relative to incongruent. We denote this Δ*Confidence.* This quantity reflects the extent to which confidence judgements are shaped by expectations. Next, we computed the BOLD difference between incongruent

Table 6.3. Results of Bayesian Confidence x Accuracy x Attention repeated-measures ANOVAS.

Bayes factors correspond to the evidence for the listed model relative to the evidence for all other models

			Bayes factors		
Region	*Confidence*	*Accuracy*	*Attention*	*Confidence + others*	*Null*
MTG	*0.14*	*0.68*	*4.96*	*0.01 - 2.26*	*1.00*
SMG	*0.85*	*1.79*	*1.96*	*0.03 - 2.67*	*1.11*
IFG (R)	*13.62*	*0.88*	*0.71*	*< .01 - 2.07*	*3.96*
AG	*1.35*	*9.83*	*0.53*	*0.01 - 3.42*	*3.47*
SG (R)	*2.64*	*0.84*	*2.42*	*< .01 - 1.13*	*10.07*
SG (L)	*1.74*	*5.52*	*1.84*	*< .01 - 1.17*	*10.22*

MTG middle temporal gyrus, SMG superior medial gyrus, IFG right inferior gyrus, SG supramarginal gyrus, AG angular gyrus and congruent reports (PE_{REPORT}), restricted to confident responses. Results showed that these quantities were negatively correlated, ρ = -.512, *p* = .027 (fig. 6.5C), confirming our finding that high confidence is associated with low PE_{REPORT} in rIFG: the more expectation increased confidence behaviourally, the more confidence was associated with low rIFG PE_{REPORT}.

To ensure that these differences were not driven by differences in reaction speed or Gabor contrast, we extracted data from the cluster revealed by our control GLM. This revealed that even after controlling for these possible confounds, rIFG BOLD was significantly higher for guess that confident responses *t*(18) = 2.21, *p* = .041, d_z = 0.44. The significant brain-behaviour correlation was also replicated, rho = -.575, *p* = .014.

Together, these analyses reveal that subjective confidence is reliably associated with PE_{REPORT} in right IFG, even after controlling for attention, Gabor contrast, decision accuracy and reaction speed.

6.3.7 Sources of priors and sensory signals for confidence.

We have shown that rIFG activity associates response prediction error with confidence. Assuming a model in which decision confidence is a weighted function of top-down expectations and 'bottom-up' sensory signals (or decision evidence), we asked whether we could identify sources of these variables. To do this we ran a seed-to-voxel psychophysiological interaction analysis (PPI), with rIFG as a functionally defined seed.

We were interested in regions communicating predictive information, and therefore regions of interest would demonstrate functional connectivity with rIFG that differs for congruent and incongruent reports. Furthermore, we wanted to determine the source of information that is used to shape confidence judgements. This can be captured by searching for regions whose congruence-dependent connectivity with rIFG is predicted by the effect of expectations on confidence. We hypothesised that expectations would be represented in frontal regions and sensory signals would be represented in visual cortex, and both of these information sources would be communicated to rIFG.

To test this hypothesis we used a behavioural covariate of interest – the influence of expectations on metacognitive bias. As in section 6.3.2, we took as our measure the extent to which expectations induce a liberal shift in confidence thresholds (C'_1+ and C'_1-) over the type 1 axis. Specifically, we computed the (mean-centred) difference between metacognitive bias for congruent and incongruent reports. This measure quantifies the reduction in type 1 evidence required to report congruent (versus incongruent) decisions with high confidence. It is independent of decision bias and accuracy. We denote this variable ΔC_1. Higher values signify that expectations liberalised metacognitive bias more.

Sources of predictive information for confidence were identified by computing the contrast incongruent ≠ congruent, with ΔC_1 as a between-subjects covariate of interest. As shown in table 6.4 and figure 6.6, the PPI analysis revealed four significant clusters: bilateral frontal pole (FP), right orbitofrontal cortex (rOFC), and right occipital pole (rOP). Increased functional connectivity between each of these regions and rIFG was associated with a stronger effect of expectations on metacognitive bias. Note that each correlation remained significant after excluding the four participants scoring highest on $\Delta C_{1.}$ Next we determined what the role of these regions might be. We extracted median beta weights from each region and subjected them to separate repeated-measures ANOVAs.

Table 6.4 PPI results.

Regions exhibiting a significant congruence x ΔC_1 interaction

Region	*Side*	*Volume (cm^3)*	p_{FDR}	*Peak MNI*		
				X	*Y*	*Z*
Frontal pole	*R*	*6.76*	*< .001*	*46*	*40*	*-20*
Orbitofrontal cortex	*R*	*5.64*	*< .001*	*20*	*26*	*-26*
Frontal pole	*L*	*3.58*	*.002*	*-40*	*44*	*-22*
Occipital pole	*R*	*3.05*	*.004*	*2*	*-94*	*20*

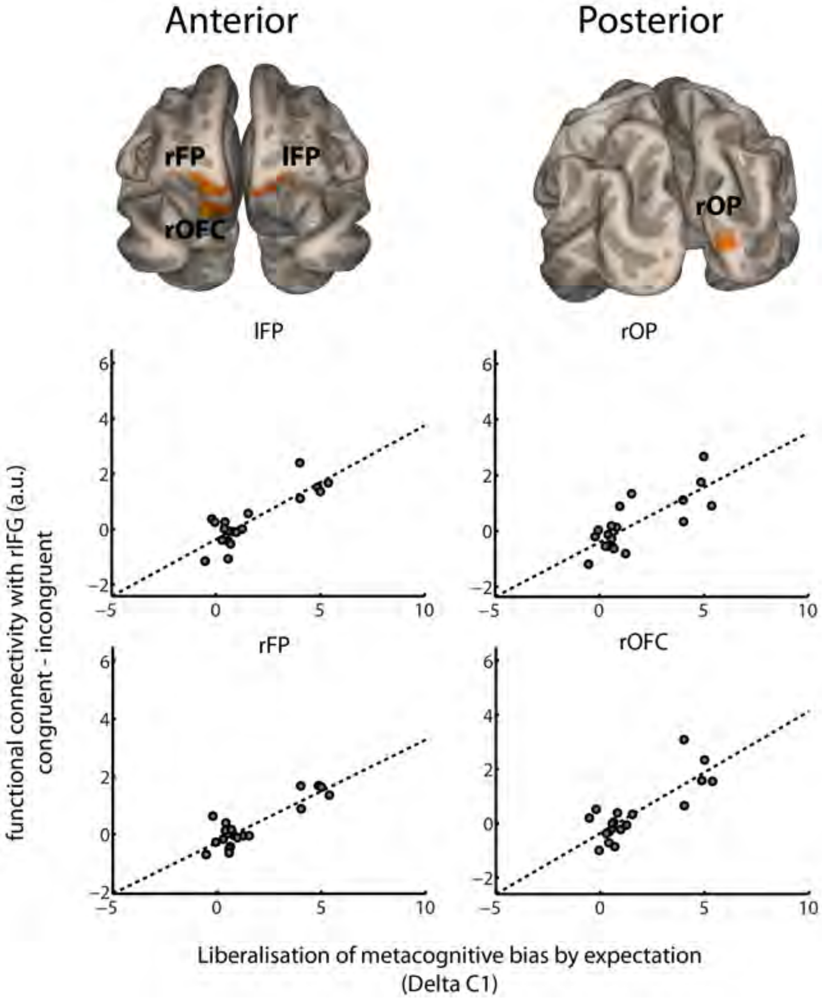

Figure 6.6 Occipito-frontal network revealed by PPI analysis

Regions revealed by PPI analysis. These are bilateral frontal pole (FP), right orbitofrontal cortex (rOFC) and right occipital pole (rOP). Each scatterplot depicts the brain-functional connectivity (FC) relationship for each region. Clockwise, these are lFP, rOP, rOFC and rFP. On the x-axis is the behavioural covariate of interest, ΔC_1. On the y-axis is the difference in FC with rIFG and each region between congruent and incongruent responses. Each scatterplot shows that increasing connectivity with rIFG is associated increasing liberalisation of metacognitive bias by expectation.

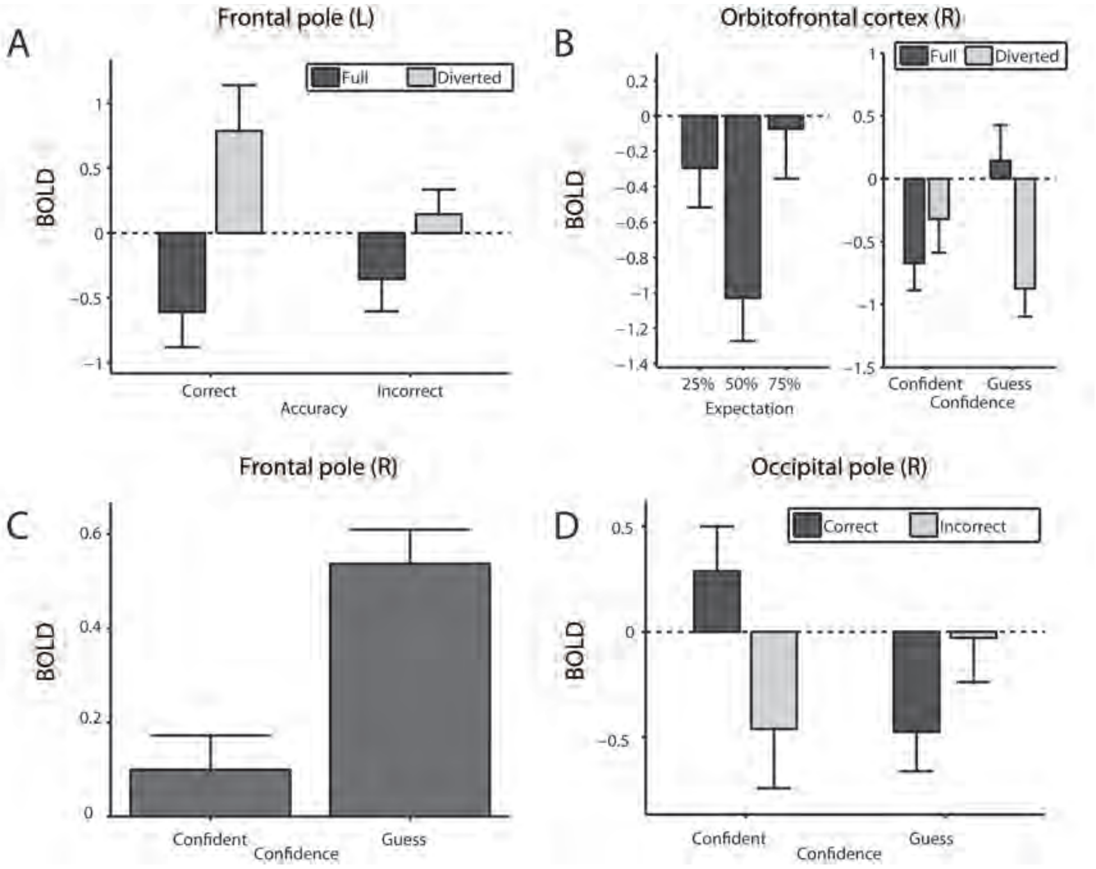

Figure 6.7. Analyses on regions in occipito-frontal network

(A) Significant attention by accuracy interaction in lFP. Correct but not incorrect decisions are associated with sensitivity to attentional state. **(B)** Significant effect of attention (left) and confidence by attention interaction (right) in rOFC. The left panel shows rOFC represents prior information. The right panel shows that attention reverses the relationship between confidence and BOLD. **(C)** Significantly greater rFP BOLD on guess than confident trias. **(D)** Significant confidence by accuracy interaction in right occipital pole. When confident, occipital pole BOLD is higher on correct than incorrect trials. There is no significant difference on guess trials. Error bars represent within-subjects SEM.

In left frontal pole, attention and accuracy interacted, *F*(1,18) = 5.99, *p* = .025, η_p^2 = .250, such that correct decisions were associated with sensitivity to attentional state, *t*(18) = 2.34, *p* =.031, but erroneous decisions were not, *t*(18) = 1.27, *p* = .221 (fig. 6.7A). Thus, decision accuracy was predicted by this region's sensitivity to attention.

In right OFC, expectation demonstrated a 'U'-shaped relationship with BOLD, *F*(1,18) = 5.33, *p* = .033, η_p^2 = .228, meaning that this region represented the prior (fig. 6.7B, left). This follows because BOLD is higher when there is an expectation (25% and 75% conditions) than in the control condition (50%). In this region there was also an attention by confidence interaction, *F*(1,18) = 7.67, *p* = .013, η_p^2 = .299, such that attention reversed the BOLD response to confidence (fig. 6.7B, right). BOLD was higher on confident than guess trials under full attention, *t*(18) = .260, *p* =.018. This pattern was reversed under diverted attention, but did not reach significance, *t*(18) = 1.82, *p* = .085. This pattern is consistent with what would be expected from (reverse) uncertainty associated with attentional state. Finally, we found a significant attention by accuracy attention, *F*(1,18) = 6.72, *p* = .018, η_p^2 = .272, whereby BOLD was higher for erroneous decisions under full, *t*(18) = 2.27, 95%CI [1.57, 2.28], *p* = .035, but not diverted, *t*(18) = 2.27, 95%CI [1.57, 2.28], *p* = .314, attention. Thus, rOFC represented prior information, attentional state and decision error.

As shown in figure 6.7C, right FP represented confidence, such that high confidence was associated with lower BOLD amplitude than guess responses, *F*(1,18) = 6,63, *p* = .019, η_p^2 = .269. Finally, in occipital pole, stimulus presentation and confidence interacted. Here, confident percepts were associated with a BOLD response that distinguished between correct and incorrect judgements, *t*(18) = 2.96, *p* = .008, whereas this was not the case for guess responses, *t*(18) = 1.11, *p* =.281. Thus, this region represents signals for perceptual sensitivity.

One might wonder whether the frontal areas directly signal priors to occipital lobe, or vice versa for sensory signals. This was not the case. Re-running the PPI analysis in the same way, but with each cluster as our seed revealed no significant connectivity with any other. We also re-ran the analysis using the change in raw confidence by expectation-response congruence as our behavioural covariate. Results were broadly consistent with those in the present analysis, revealing an occipito-frontal network recruited when expectations are integrated into subjective judgements. Specifically, we found a significant cluster in right orbitofrontal cortex (MNI *x* = 10, *y* = 28, *z* = -18, p_{FDR} = .024), left orbitofrontal cortex (MNI *x* = -36, *y* = 38, *z* = -18, p_{FDR} = .008) and in intracalcarine sulcus (MNI *x* = 6, *y* = -58, *z* = 12, p_{FDR} = .004).

Taken together, these results show that the integration of expectations into confidence judgements recruits an occipito-frontal network that represents top-down influences of attention and expectation in frontal regions, and decision-related signals in sensory cortex.

6.3.7 The contribution of visual regions and OFC to confidence is predicted by white matter density

Our connectivity analyses revealed that OFC/FP and visual cortex represented top-down and bottom-up signals respectively, and that the recruitment of these regions was predicted by the effect of confidence on metacognitive bias. The presence of these individual differences motivated an exploratory follow-up analysis that asked whether they are reflected in brain structure. More specifically, we considered whether the weighting of top-down predictions and bottom-up signals was a function of white or grey matter (WM and GM respectively) density of the source regions.

The BOLD response of our cluster in OFC reflected an effect of perceptual expectations on objective decision. The behavioural correlate of this is therefore $\Delta\ c = c_{25\%} - c_{75\%}$, - the extent to which perceptual expectations bias (yes/no) decision. We performed a whole-brain multiple regression analysis on WM density, with total intracranial volume and participant age as nuisance covariates, and with Δc as the regressor of interest. This analysis revealed that propensity to incorporate low-level priors into decision-making, as measured by Δc, was negatively correlated with rOFC white matter density (fig 6.8A and B, peak MNI $x = 23$, $y = 30$, $z = -14$, 11.51cm^3, $P_{\text{peak-FWE}} = .030$, Z = 5.08). The same analysis for GM yielded no significant results.

Given that both rOFC and rIFG BOLD predicted confidence we performed the same analysis, but this time with mean confidence as the regressor of interest. Mean confidence represents one’s overall belief in their perceptual performance, or self-efficacy. Higher values correspond to higher confidence in one’s decision-making ability (versus trial-by-trial performance). This revealed a significant cluster in contralateral occipital lobe. Here, increasing WM density

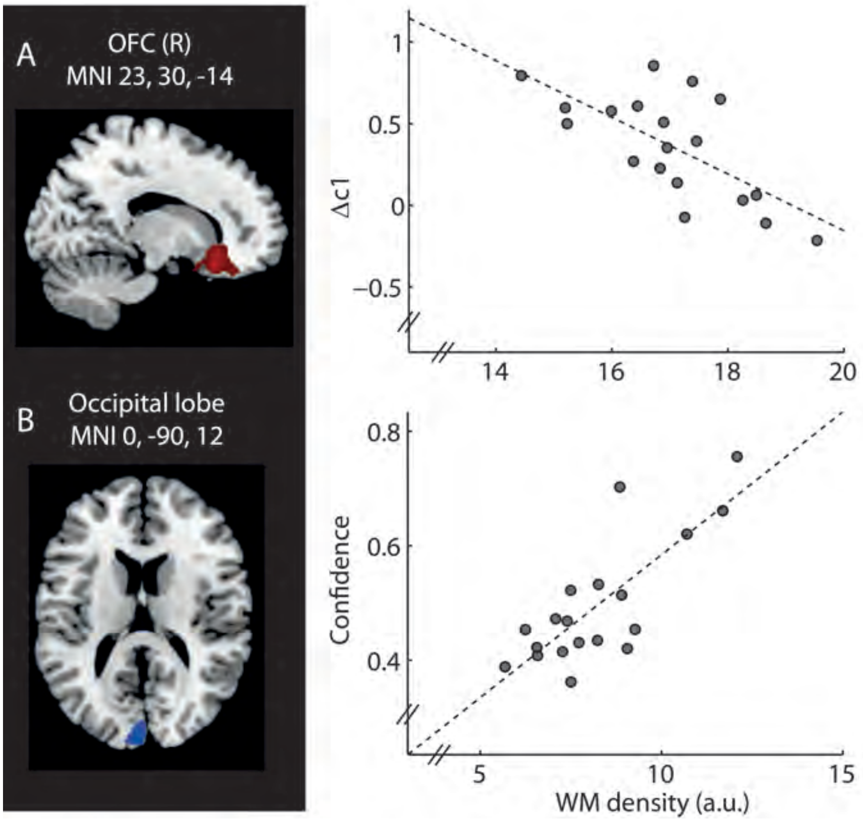

Figure 6.8. VBM results.

A. White matter density in right orbitofrontal cortex negatively predicts the effect of expectation on perceptual decision. **B.** White matter density in contralateral occipital pole is positively correlated with mean confidence across trials.

significantly predicted greater confidence at the cluster, but not the peak level (figure 6.6C and D peak MNI x = 0, y = -87, z = 16, 6.40cm^3, $p_{\text{peak-FWE}}$ = .789, p_{FWE} = .028).

Together these results suggest that the dependence of confidence on functional connectivity with source regions is reflected in anatomical indications of that connectivity: WM density in OFC was negatively predicted by its functional correlate; and increasing occipital pole WM density was associated with mean confidence, that is, beliefs of better perceptual performance.

6.4 Discussion

In the present paper we have shown that behavioural confidence in perceptual decision increases when decisions are supported by (or congruent with) prior expectations. Crucially, we show that this predictive information is, at least in part, integrated into confidence in right inferior frontal gyrus (rIFG).

We have shown that unexpected percepts, taken with respect to the decision or report, are associated with a heightened BOLD response (termed here PE_{REPORT}) in a distributed set of frontal, parietal and temporal decision-related regions. Interestingly, this expectation-sensitive set resembles those implicated in other forms of 'top-down' processing such as modality-independent sensory change detection (Downar, Crawley, Mikulis, & Davis, 2000), response inhibition (Criaud & Boulinguez, 2013; Verbruggen & Logan, 2008), and detection of behavioural salience (Jonathan Downar, Crawley, Mikulis, & Davis, 2002).

Our crucial result was that the contribution of top-down expectations to subjective confidence judgements was reflected in fMRI BOLD, specifically in right inferior frontal gyrus (rIFG). Here, high confidence was associated with a lower prediction error response profile. Furthermore, the more that confidence was shaped by expectation behaviourally, the more that confidence was associated with low prediction error signals in this area. Our results therefore indicate a central role for rIFG in perceptual decision making in which the 'match' between internal templates and perceptual content is integrated into subjective confidence judgements.

Under an alternative account, the sensitivity of rIFG to confidence would be an indirect effect of sensitivity to task difficulty. For example, rIFG may infer task difficulty from the degree to which the percept is surprising. However, this interpretation was ruled out by control analyses, which showed that the PE_{REPORT}-confidence relationship was not driven by choice accuracy. These control analyses additionally excluded attention, stimulus contrast, and reaction speed as driving the observed relationship between PE_{REPORT} and confidence in rIFG.

This process of relating predictive information into confidence judgements recruited both occipital lobe and frontal regions bilateral frontal pole (FP) and right orbitofrontal cortex (rOFC). In left occipital lobe, decision-related signals were represented. Interestingly, connectivity between rIFG and contralateral occipital lobe was not found. One possibility is that contralateral occipital lobe is functionally connected with rIFG independently of expectation effects; another possibility is that statistical power was too low to detect connectivity reflecting the neural response to a small stimulus in retinotopically-organised space. We interpret the functional connectivity with occipital lobe as the communication of sensory signals. By contrast, we found the representation of top-down influences in frontal regions. In particular, right OFC represented prior information, consistent with previous work (Schoenbaum & Roesch, 2005; Trapp & Bar, 2015; Wallis, 2007), and white matter density in this area even predicted behavioural effects of expectation on objective decision. Right OFC was also sensitive to attentional state. Here, representation of the prior required attention, and furthermore, the BOLD response to decision confidence reversed with attention. Under full attention rOFC BOLD was higher for guess responses than confident responses, as is usually found (Fleming et al., 2012; Hilgenstock, Weiss, & Witte, 2014). However, under diverted attention this pattern reversed, possibly indicating that rOFC represents the uncertainty associated with attentional state: high under full attention, but low under diverted attention.

Altogether, we interpret these results as showing that subjective confidence is represented in rIFG as a combination of both stimulus-driven signals, communicated from occipital lobe, and shaped by top-down perceptual expectations and attention, communicated from OFC and frontal pole. OFC has been repeatedly been shown to reflect reward expectations and beliefs (De Martino et al., 2013; Kepecs et al., 2008; Kim, Shimojo, & O’Doherty, 2011; Lebreton et al., 2015), however here we place OFC belief representations within a larger hierarchical structure for perceptual processing, generating predictions (Stalnaker, Cooch, & Schoenbaum, 2015; Trapp & Bar, 2015) that constrain subjective confidence judgements in perceptual decision.

Importantly, our PPI analysis cannot determine the directionality of functional connections in this network. One possibility is that rIFG is involved in constructing confidence from an integration of PE_{REPORT} signals and top-down expectations. Here our occipito-frontal network would be sending signals to rIFG. However, another possibility is that bottom-up signals are passed from occipital lobe to rIFG, and an initial transformation of $\text{PE}_{\text{REPORT}}$ into confidence is signalled to the frontal regions of this network. Under such an account, the role of rOFC and/or frontal pole may be one which transforms the confidence estimate represented in rIFG into a reportable judgement, based on the mismatch between the estimate, expectations, and potentially, attentional state (Lebreton et al., 2015). Further studies will be needed to disambiguate these possibilities.

Our results are readily interpretable from Bayesian brain perspectives (Clark, 2013; Friston, 2009; Lee, 2002; Yuille & Kersten, 2006b). These propose that perceptual inference is a weighted integration of sensory evidence and prior beliefs about the cause of the sensation, such that the perceptual report corresponds to the belief with the greatest posterior probability. The posterior probability increases as the correspondence between prior and sensory signal increases. Therefore, inference is deemed ‘successful’, and so should be associated with high confidence, when we see a low ‘prediction error‘ response, as we saw here (Meyniel, Sigman, et al., 2015). Neuronal representations of prediction errors are well-established in the reward domain (Bayer & Glimcher, 2005; Nakahara et al., 2004), but in the perceptual domain evidence remains restricted to BOLD correlates such as $\text{PE}_{\text{REPORT}}$. Under such a Bayesian brain account, our connectivity results suggest that occipital lobe sulcus passes sensory signals to rIFG, and frontal pole/OFC passes top-down predictions and weightings of attentional state. In this view, the finding that $\text{PE}_{\text{REPORT}}$ amplitude in rIFG was lower for confident responses is consistent with the representation or construction of the posterior belief in this region. This in turn is in line with empirical evidence for rIFG encoding of the decision variable, either in Bayesian form (the posterior; d’Acremont et al. 2013) or as decision evidence (Hebart et al., 2014), which are mathematically equivalent, (Bitzer et al. 2014).

Previous work has separately implicated rIFG in the representation of both the decision variable (Bubic et al., 2009; d’Acremont et al., 2013) and expectation violation in a range of modalities, from speech perception (Clos et al., 2014) to the auditory (Garrido, Kilner, Kiebel, & Friston, 2009), visual (Bubic et al., 2009), and tactile (Allen et al., 2016) domains. Previous work has also implicated rIFG in the representation of subjective uncertainty (Fleck, Daselaar, Dobbins, & Cabeza, 2006; Fleming & Dolan, 2012). However, to our knowledge these functions of rIFG have not been related to each other before. Right IFG has also been implicated in a wide range of related executive processes such as novelty detection (Hampshire, Chamberlain, Monti, Duncan, & Owen, 2010), change detection (Beck, Rees, Frith, & Lavie, 2001), and behavioural relevance (Hampshire et al., 2010). Furthermore, it has been implicated in detecting or resolving response conflict (Casey et al., 2000; Hampshire et al., 2010), and is a key component of the response inhibition network (Criaud & Boulinguez, 2013; Verbruggen & Logan, 2008). This raises the intriguing possibility of a functional overlap between resolution of response conflict and the formation of confidence.

These roles could be unified by considering rIFG as the region in which the posterior is computed (at least for perceptual tasks), because the posterior belief on sensory causes affords a hypothesis space for adaptive, plausible actions (Mansouri, Tanaka, & Buckley, 2009). Such a view is consistent with evidence for rIFG in appropriately acting on perceptual choices (Suzuki & Gottlieb, 2013), computing behavioural significance (Sakagami & Pan, 2007) , computing action-outcome likelihoods that modulate motor cortex (Morris, Dezfouli, Griffiths, & Balleine, 2014), and representing the posterior (d'Acremont et al., 2013). It has even been shown that the rIFG BOLD response to decision errors is associated with both the valence of the decision outcome, and the optimism of the participant ('self-belief'; Sharot et al. 2011), consistent with a view of rIFG in which high-level, abstracted posteriors are computed from beliefs and errors. Anatomical considerations support such a view, since the rIFG is directly connected with regions relevant for both cognitive and motor control (Petrides & Pandya, 2002). We leave open for future research the question of whether and how rIFG relates perceptual confidence to action outcomes.

6.5 SUMMARY

In summary, we have shown that top-down expectations are integrated into decision confidence, and have shown that this occurs in a functional network consisting of rIFG, bilateral frontal pole, right OFC and occipital lobe. Here, top-down perceptual expectations and bottom-up sensory inputs are integrated into a subjective sense of perceptual confidence. Together, our data reveal a crucial role of top-down influences in the mechanism by which perceptual decisions become available for conscious report.

7

GENERAL DISCUSSION

7.1 Overview

Subjective perceptual confidence is an integral component of visual consciousness, which in part reflects the strength of perceptual experience (Fleming & Lau, 2014). As such, understanding how confidence is constructed may help us to get closer to understanding the mechanisms underlying subjective experience. In parallel, by understanding how the correspondence between confidence and accuracy arises we can better understand the mechanisms underlying our ability to introspect upon and evaluate our decisions.

A large portion of the literature on perceptual confidence considers how sensory information shapes these judgements, and the conditions under which confidence and accuracy dissociate. However, the contribution of top-down influences to the formation of subjective judgements has remained understudied. This is surprising, despite the fact that their contribution to the formation of objective decisions are of great interest. To this end, the work in this thesis has investigated how perceptual confidence is shaped by top-down perceptual expectations, or priors. A novel behavioural paradigm has been introduced, which consistently shows that confidence and metacognitive bias increase, and metacognition improves, for perceptual decisions that have a high prior probability, that is, are expectation-congruent. Using EEG and fMRI, Chapters 5 and 6 examined how these priors are integrated into confidence before and after target onset. This discussion chapter will consolidate the empirical findings presented, and suggest a plausible extension of Bayesian brain frameworks into the domain of subjective reports.

7.2 Key results

7.2.1 Subjective judgements are shaped by expectations, not attention.

Interest in how expectations shape visual consciousness is growing, and we now know that percepts receive preferential access to awareness when they are predicted in either content (Chang et al., 2015; Melloni et al., 2011; Pinto et al., 2015; Sterzer et al., 2008) or time (Mathewson et al., 2012; Wyart & Tallon-Baudry, 2009). These effects may arise at the level of objective perception, improving the fidelity of the perceptual representation itself. Alternatively (or additionally) effects of expectation could reflect changes in the threshold for awareness, such that more probable percepts are advantaged in their access to visual consciousness. Finally, given that attended objects are more likely to reach awareness (Kanai, Tsuchiya, & Verstraten, 2006; Kanai et al., 2010; Lavie, 2006; Wyart & Tallon-Baudry, 2008), top-down influences of expectation may simply reflect attentional effects. The work presented here has investigated subjective confidence judgements while factorially manipulating attention and expectation.

Work here has shown that perceptual decisions that are more *a priori* probable, that is, are expectation-congruent, are associated with both liberalised metacognitive bias (confidence) and improved metacognitive accuracy. While attention exerts a strong effect on detection sensitivity (as measured by contrast thresholds), its relationship with subjective measures is less clear. Metacognitive bias, that is, the tendency to report decisions with high confidence (independently of accuracy) became more liberal with expectation independently of attention. However the relationship between attention, expectation and the proportion of confident trials was less stable. It may be that the dependence of expectation on attention depends on stimulus contrast. As shown in table 7.1, Chapters in which the effect of attention on contrast sensitivity was stronger revealed a dependence of expectation effects on attention. Such an explanation for the relationship between attention and

Table 7.1 Interactions between attention and expectation by Chapter

	Effect of inattention on expectation-confidence relationship	*Stimulus contrast (attended vs. unattended)*
Chapter 4	*Eliminates relationship*	*8% vs. 3% (62% change)*
Chapter 5	*Dampens relationship*	*19% vs. 25% (24% change)*
Chapter 6	*Relationship still present*	*5% vs. 4% (20% change)*

contrast would be consistent with recent work, which shows subjective visibility ratings of unattended targets to be largely insensitive to signal strength (Rahnev et al., 2011). Rahnev and colleagues' data reveal that that as signal strength increases, attended targets will (unsurprisingly) be associated with higher confidence, yet confidence for unattended targets remain unaffected.

We therefore see an effect of attention on confidence that increases in magnitude with stimulus contrast. Under such an account, expectations do not require attention to shape confidence as such. Rather, interactions between attention and expectation are driven indirectly, by failures to incorporate signal strength into confidence judgements under inattention.

Consistent with this interpretation, the neural mechanisms underlying expectancy effects on confidence were largely independent of attention. Pre-stimulus alpha oscillations, previously implicated in top-down attentional effects on sensitivity (Busch & VanRullen, 2010; Landau & Fries, 2012; Mazaheri, DiQuattro, Bengson, & Geng, 2011; Rohenkohl & Nobre, 2011; Zumer et al., 2014), are here shown to be involved in effects of expectation, not attention. The phasic modulation of both objective and subjective perception by alpha oscillations persisted when attention was diverted from the task. Similarly, the neural response to violations of expectations – perceptual 'prediction error' – and its negative relationship with confidence is independent of attention.

This does not mean that confidence is wholly independent of attention. For example, results showed that right orbitofrontal cortex exhibited opposing responses to confident versus uncertain decisions, depending on attentional state. Neither do these results mean that attention and expectation do not interact. Indeed, Chapter 4 shows that perceptual sensitivity for task-relevant decisions is increased when perceptual decisions on a secondary task are shaped by valid priors, suggesting that correctly predicting unattended perceptual content reduces the associated processing load and frees up resources for primary tasks (Hohwy, 2012; Sy, Guerin, Stegman, & Giesbrecht, 2014).

These results implicate top-down attention primarily in shaping objective perception, whereas the effects of top-down expectation appear to permeate into the domain of subjective perception as well. To the extent that perceptual metacognition can be used as a proxy for visual awareness (Dienes & Seth, 2010a; Fernandez-Duque et al., 2000; Kentridge & Heywood, 2000; Kunimoto et al., 2001; Sandberg et al., 2010; Seth et al., 2008, but see Jachs et al., 2015), these results suggest that top-down attention may act at the level of sensory representation, while expectations bias objective and subjective thresholds so that probable perceptual inferences receive preferential access to visual consciousness.

These conclusions about expectation and attention are more directly applicable to metacognition. The data presented here consistently reveal an interaction between expectation and perceptual decision, indicating that metacognition arises, at least in part, from some comparison process between the perceptual decision and its prior probability, akin to what is thought to occur for objective perception (Summerfield & de Lange, 2014), and consistent with evidence for confidence tracking posterior probabilities (Aitchison et al., 2015; Feldman & Friston, 2010; Meyniel, Schlunegger, et al., 2015). While the effects of expectation on subjective judgements are somewhat dampened under diverted attention, attention does not appear to be necessary for appropriately placing confidence thresholds.

It is important to note that, while not the focus of this thesis, strong conclusions about whether metacognitive accuracy increases with expectations cannot be drawn. When estimating metacognition separately for yes and no reports, meta-*d’* is neither robust to changes in metacognitive bias (confidence) nor to changes in decision thresholds (Barrett et al., 2013; Evans & Azzopardi, 2007; Fleming & Lau, 2014). Type 2 ROC curves, though invariant to changes in metacognitive bias by design, are also inappropriate for the data presented here because they are biased by decision thresholds (Galvin et al., 2003).

The effect of expectation-response congruence on confidence is easily accommodated by normative perceptual decision-making models. In evidence accumulation terms, the prior on some perceptual choice is generally modelled as its initial evidence, C_0.(Dunovan, Tremel, & Wheeler, 2014; Wyart et al., 2012). This means that at point of decision, evidence for probable percepts will be higher if the initial evidence had favoured the unselected choice. To the extent that confidence can be formulated as decision evidence, expectation-congruent reports will therefore be associated with higher confidence than their incongruent counterparts (Ratcliff & McKoon, 2008; Vickers, 1970; Wong, 2006). This account is easily reformulated in Bayesian terms. To the extent that confidence can be modelled as a function of the posterior belief (Feldman & Friston, 2010; Hangya, Sanders, & Kepecs, 2016; Hebart et al., 2014; Meyniel, Sigman, et al., 2015; Pouget et al., 2016), confidence will increase with increasing prior probability of that decision. On both accounts, the behavioural and pre-stimulus effects of expectation on confidence presented in this thesis can be explained in terms of by biases at the beginning of the evidence accumulation process. However, this account cannot explain why confidence thresholds over the type 1 axis also changed with expectation. While further work is needed to assess the behaviour of this confidence measure in depth, results here suggest that the threshold for reporting decisions with confidence changed over and above changes in type 1 criterion.

The extent to which this account can explain effects of congruence on metacognition is unclear. The analyses here largely used the measure *meta-d*'/*d'* to measure metacognitive accuracy, which should be invariant to differences in *d',* decision threshold *c*, and metacognitive bias *C* (Barrett et al., 2013; Maniscalco & Lau, 2012). In theory, this should mean that effects of expectation on metacognition cannot be explained by changes in confidence thresholds alone: *meta-d'/d'* should be invariant to decision and confidence biases. One issue is that when *meta-d'* is estimated separately for “yes” and “no” judgements, as was necessary here, the measure loses its invariance to bias (Barrett et al., 2013). Therefore the results found under metacognitive accuracy might in fact be biased by changes in decision and confidence thresholds. Indeed, this is how the results were modelled in Chapter 4.

A second possible issue is that expectations might have changed decision accuracy, despite seeing no differences in *d'.* By design, expectation-congruent decisions are more likely to be correct because the expectation is valid. To illustrate, ‘yes’ responses have an *a priori* 75% chance of being correct in the ‘expect present’ condition but only a 25% chance of being correct in the ‘expect absent condition’. While meta-*d’* can be computed separately for yes and no responses, *d'* cannot. This means that response-conditional meta-*d’*/*d’* values in these studies may also be dependent on accuracy, because *d’* cannot capture sensitivity on congruent versus incongruent trials. Future research could investigate these possibilities in a number of ways. The most straightforward approach would be to use a paradigm that manipulates expectation without relying on response-conditional measures, for example, by measuring metacognition while environmental statistics are learned. Alternatively, false perceptual beliefs could be induced. This latter manipulation should decouple accuracy from predictability, because expected percepts would be not be more likely to be correct.

7.2.2 Frontal and sensory contributions to the integration of top-down influences on confidence

Chapters 5 and 6 investigated the neural mechanisms that underlie top-down effects on confidence. Chapter 5 showed that the influence of perceptual priors can be predicted by ongoing brain activity, such that sorting trials according to stages of the pre-stimulus alpha cycle reveals certain phases that are associated with stronger effects of expectation on both decision and on confidence. The phases at which decision thresholds were most biased by expectations were also those at which confidence thresholds were most biased by expectation-response congruence. This suggests that at the decision stage, confidence judgements incorporated the expectancy information that had been made available pre-stimulus: stronger prior evidence for the decision increased confidence, whereas stronger prior evidence against the decision decreased confidence.

In turn, these results show that the propensity to integrate priors into decisions is dependent on fluctuations in cortical excitability over visual regions (Lindsley, 1952). Expectations maximally bias decisions by expectation every 100ms, and these 10Hz cycles in which top-down (versus bottom-up) signals dominate objective decision-making support the view that oscillations may carry prior information to task-relevant brain areas (Bastos et al., 2012; Engel et al., 2001; Friston, 2012; van Kerkoerle et al., 2014; von Stein et al., 2000).

So, what process is reflected in the alpha cycle? One possibility is that alpha oscillations reflect the recruitment of prior evidence to visual regions. Such an account is consistent with proposed role of alpha oscillations in long-range communication across cortical areas and in top-down control (Arnal & Giraud, 2012; Engel et al., 2001; Fries, 2005; Palva & Palva, 2007; Womelsdorf & Fries, 2007). From an evidence accumulation stand-point, this could correspond to a shift in baseline evidence, such that the relevant neural populations need less sensory evidence to fire (Summerfield & de Lange, 2014). In Bayesian terms, this could correspond to fluctuations in the mean of the prior belief, itself leading to fluctuations in the posterior. On an alternative view of the alpha cycle, decision threshold fluctuations may correspond not to the availability of prior evidence, but to the weighting of prior evidence. Under this account, the prior probability would be constant with respect to alpha phase, however sensory precision would be subject to 100ms cycles. Phases at which sensory signals are represented with higher fidelity may be associated with a reduced effect of expectation. This model would be consistent with previous evidence for pre-stimulus phase- modulation of visual attention using spatial cueing paradigms (Busch & VanRullen, 2010; Frey, Ruhnau, & Weisz, 2015; Landau & Fries, 2012).

Under both accounts there exist optimal levels of cortical excitability for the incorporation of priors into decision. Both accounts also suggest that expectations modulate cortical excitability in order to adaptively facilitate or inhibit neural responses to forthcoming signals, and both accounts propose the existence of an optimal level of cortical excitability that depends upon whether stimulus presentation is expected or not. However the data presented here cannot distinguish between the fluctuation-in-mean and fluctuation-in-precision explanations. Future work could address this question by, for example, using model-based EEG to estimate trial-by-trial values of both mean and weighting (precision, e.g. modelling the data under the hierarchical Gaussian Filter, Mathys et al. 2014). Alternatively, these two variables could be manipulated orthogonally in the direction of dots presented in a random dot kinematogram (RDK), so that participants learn to expect a particular distribution of dot motion. If alpha phase reflects precision then it should predict the influence of expected variance (sensitivity), whereas if alpha phase reflects the prior then it should predict the influence of expected mean (bias).

Using fMRI, Chapter 6 revealed a range of cortical regions that are sensitive to the mismatch between percept (decision) and prior, and one of these regions – right inferior frontal gyrus (rIFG) – represented a confidence signal that was dependent upon this signal. Here, lower mismatch responses (‘prediction error’) predicted higher confidence. This results is consistent with a role of rIFG in non-spatially re-orienteering attention to targets whose improbability signals behavioural relevance (Corbetta & Shulman, 2002), and in integrating sensory and motivational information to drive goal-directed behaviour (Sakagami & Pan, 2007). This process of integrating priors and sensory signals into confidence recruits occipital lobe, the source of bottom-up signals, bilateral frontal pole (FP), representing confidence and attentional state, and right orbitofrontal cortex (OFC), which represented both priors and representations of attentional state. Specifically, OFC activity for confident versus guess responses reversed under diverted attention, so that BOLD was higher for guess responses under full attention, but higher for guess responses otherwise. These results suggest that OFC may track the uncertainty arising with attentional state, either communicating this to rIFG or shaping confidence signals from rIFG.

Do these findings – of pre-stimulus modulation by expectations in occipital areas, versus the representation of expectations in rIFG and OFC in the post-stimulus period – conflict? One might imagine that OFC should have shown functional connectivity with occipital lobe, reflecting the ongoing (i.e. pre-stimulus) communication of priors that periodically shape decision-making. These results are interpreted as reflecting distinct processing stages. While the EEG results of Chapter 5 implicate priors in periodically altering baseline evidence for probable decisions, the fMRI results of Chapter 6 implicate OFC in communicating priors that are matched against the sensory signals (in sensory regions) by rIFG. Dunovan and colleagues have shown that priors are indeed incorporated into decision-making at two stages. First, baseline evidence is set according to perceptual priors, and weighted by the reliability of that prior. Secondly, evidence accumulation rate is determined dynamically, according to the correspondence between evidence and prior (Dunovan et al., 2014). Together, these results suggest that initial evidence may indeed be communicated to visual areas, possibly by OFC, prior to target onset, but following target onset, priors are continually compared against sensory evidence in rIFG, with evidence being accumulated faster for high match trials, i.e., expectation-congruent responses.

7.3 Confidence in the Bayesian brain: a framework

How might perceptual priors shape confidence? The work in this thesis suggests that perceptual expectations bias the evidence for the chosen sensory hypothesis in favour of more probable sensory hypotheses, and that this is instantiated by both sensory and frontal regions. These results support the widely held view that decision and confidence are based, at least in part, upon a common evidence source (Kepecs & Mainen, 2012; Kiani et al., 2014; Ratcliff & Starns, 2013). However, the work in Chapter 6 suggests that the evidence source that is relevant for confidence is not just sensory information, but also a re-representation of the decision itself (Cleeremans, 2011): the discrepancy between expectation and choice.

Computational models of decision-making make the intuitive proposition that perceptual choices will correspond to the option with the most decision evidence, the ‘balance of evidence hypothesis’. In Bayesian terms, this proposition corresponds to the sensory cause associated with the peak of the posterior probability distribution, that is, the belief with the greatest posterior probability (Meyniel, Sigman, et al., 2015; Pouget et al., 2016). On one Bayesian account of confidence, confidence is the variance, or precision of this distribution (Meyniel, Sigman, et al., 2015). This 'distributional confidence' seems to capture uncertainty rather than choice confidence. The variance of the posterior *pdf* is orthogonal to the mean, so here, confidence is not defined in terms of the decision to which it pertains, going against common conceptions of choice confidence (Kvam, Pleskac, Yu, & Busemeyer, 2015; Pouget et al., 2016).

A popular alternative proposes that confidence corresponds to the posterior probability of the decision, given the evidence. This formulation captures the definition of choice confidence well: the subjective probability of the decision having been correct. Where confidence judgements are collected on a scale, the assumption is that there exists some threshold such that if the posterior probability exceeds the threshold confidence is reported as 'high', and otherwise it is deemed to be 'low'. This process of bifurcating continuous representations of confidence onto a reportable scale has been linked to orbitofrontal cortex (Lebreton et al., 2015). But how is this threshold set? Modelling confidence in terms of the posterior belief alone does not address the question of how decision confidence is computed and made available for report.

While the perceptual decision-making literature considers confidence as a product of the objective decision-making process, other domains conceive confidence as a 'second-order' decision-making process. Here, confidence judgements are considered to be 'meta-decisions' in the sense that they are inferences on the accuracy of one's decision. Some previous attempts to explain confidence have indeed proposed a 'read-out' of first order evidence. For example, type 2 signal detection theory (Evans & Azzopardi, 2007; Galvin et al., 2003) assumes an internal representation of being *objectively* correct, while higher order thought theory posits that conscious states correspond to 'higher-level representations' of first order states (Gennaro, 1996; Lau, 2007; Rosenthal, 2000; Timmermans, Schilbach, Pasquali, & Cleeremans, 2012). The problem here is that such accounts require some monitoring or 'read-out' system, meaning that as representations become more abstracted, the system must accommodate an increasing number of monitoring layers. To illustrate, I may be confident in a choice I have made, but feel that my sense of confidence is not a good predictor of positive outcomes. While one monitoring layer (for confidence judgements) may be neurobiologically plausible, as more become necessary the plausibility of such a system decreases.

Hierarchical Bayesian frameworks circumvent this issue of requiring specialised modules for decisions at each level of abstraction. When processing is hierarchically organised, it can move into increasing levels of abstraction and re-representation without requiring any additional mechanisms, because for any decision the output of each hierarchical stage (its inference) will be a function of its inputs (top-down priors and bottom-up data). Here, there is no ‘monitoring layer’, as such, because every layer in every processing stream both receives input from subordinate levels, but also constrains levels above: *every* layer is a monitoring layer. Crucially, there is no upper-most layer, because the topology of the predictive coding hierarchy is more akin to a torus (doughnut), instantiating interdependent inferences across the brain. Thus, hierarchical predictive coding implicitly incorporates monitoring layers, but here these layers are embedded within a neurobiologically plausible system.

Higher-order decisions (‘meta-inference’) in a hierarchical Bayesian framework should be computationally and mechanistically analogous to lower-order decisions. They only require the capacity for representing the relevant prior. For example, orientation discrimination requires the relative probability of leftwards orientation in V1. However for subjective confidence judgements, the relevant priors must pertain to that confidence judgement – the prior probability of the decision having been correct.

Figure 7.1 presents a broad overview of how the construction of confidence could be achieved in a hierarchical Bayesian scheme. The proposed model assumes a predictive coding scheme, in which priors are passed via feedback connections and prediction errors are passed feed-forward. Following previous

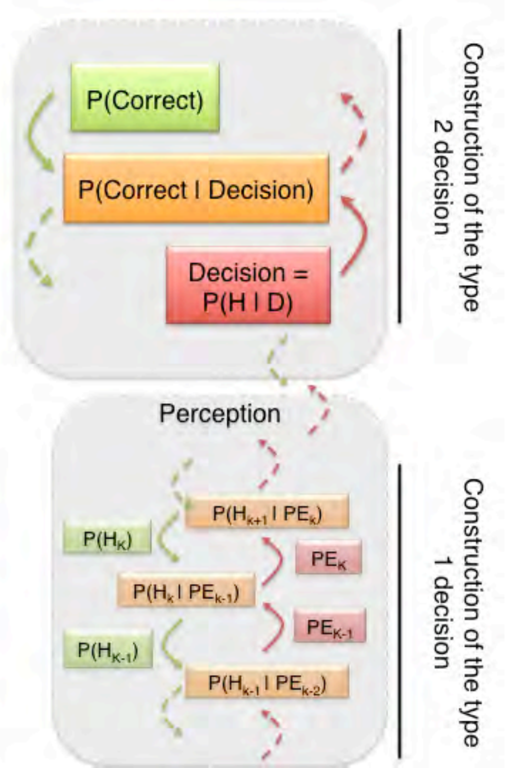

Figure 7.1. A model for subjective confidence in Bayesian schemes.

The type 1 decision is constructed in the manner predicted by hierarchical predictive coding models. At level k of the perceptual hierarchy, top-down priors are received from level k+1, and bottom-up prediction errors are received from level k-1. The inference at level k corresponds to the hypothesis with maximal posterior probability, given the prediction error. This posterior belief will form an empirical prior on the inference at level k-1. Bottom-up input to level k+1 will be the remaining prediction error. The type 2 decision is constructed from a top-down 'meta-prior' – the prior probability of making a correct report – and bottom-up prediction error, corresponding to the discrepancy between the perceptual decision (given by the posterior) and perceptual prior (prior probability of the selected sensory cause).

work, the model assumes that each level within the perceptual hierarchy receives bottom-up signals (here, prediction errors) and relevant top-down priors, and integrates them in order to identify the hypothesis with maximal posterior probability. This hypothesis (the posterior belief) becomes an empirical prior for the level below, and any remaining discrepancy between data and posterior belief becomes the prediction error for the level above. This part of the model is taken directly from previous work (Friston, 2009; Knill & Pouget, 2004; Lee & Mumford, 2003; Spratling, 2016; Yuille & Kersten, 2006).

The extension of this framework into confidence judgements is straightforward. While yes/no reports are determined according to the posterior belief of target presence, confidence in that judgement is determined according to the posterior belief of being correct. Given that the goal is to infer the accuracy of a decision, the relevant prior will be the prior probability of making a correct decision. Similarly, the bottom-up information will be the prediction error from the level below: the sensory evidence unexplained by the reportable perceptual decision.

This variable is simply a predictive coding formulation of ‘decision evidence’. So, while objective perceptual decisions pertain to the mean of the posterior belief on sensory causes, conditioned upon sensory evidence and perceptual priors, confidence will be the mean of the posterior belief on decision accuracy, conditioned upon decision evidence and expected task performance, or self-efficacy. The framework presented here bears similarity to the Radical Plasticity Thesis of Cleeremans and colleagues (Cleeremans, 2011; Timmermans et al., 2012), who propose that metacognitive processes arise from a subpersonal re-representation of lower-order states. They propose that objective decisions are determined according to the activity of a first-order layer, and its outputs form the inputs of a second layer that learns to predict the errors in the first.

It is important to note that because the empirical chapters of this thesis did not manipulate prior beliefs about performance, the data presented in this thesis cannot support or refute this model. The data in this thesis only support the notion that perceptual priors shape confidence, most likely at the level of the first-order inference (Chapters 5 and 6). Under the model proposed here, perceptual priors shape the posterior belief, and so only indirectly shape confidence. This model motivates the hypothesis that manipulating beliefs about perceptual performance, for example by giving blockwise feedback, would shape confidence more than sensitivity. It also motivates the hypothesis that trial-by-trial retrospective confidence can be modelled as an integration of prospective confidence and decision evidence.

Another way of probing beliefs about task performance is to estimate confidence thresholds, averaged over all conditions and responses. This measure represents participants' overall belief that they have given a correct response. Exploratory correlations on each node of the occipito-frontal network against mean confidence threshold (defined over the type 1 axis) revealed that left frontal pole and right orbitofrontal cortex may represent these 'self-efficacy' priors (figure 7.2), though this should be confirmed with further research.

How might these self-efficacy priors – that is, expected decision accuracy – be learned? Hierarchical predictive coding recruits the notion of 'empirical priors', where each prior is constrained by the inferences at higher hierarchical stages.

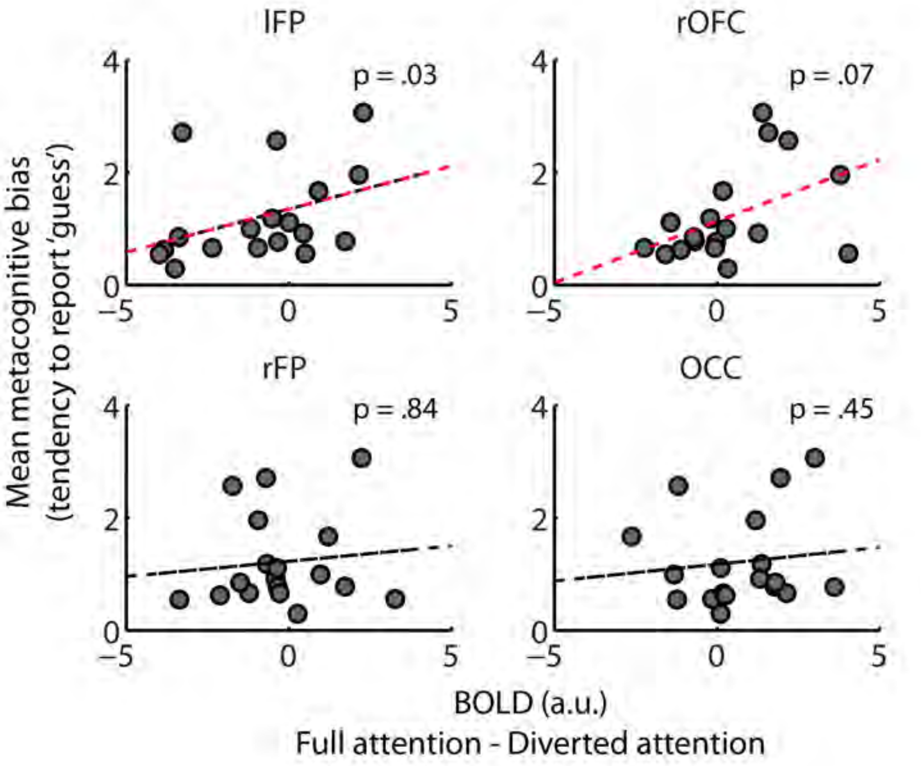

Figure 7.2 Self-efficacy (mean width of confidence thresholds) in left frontal pole.

Exploratory correlations showing the relationship between self-efficacy, as defined by overall confidence bias (width of confidence thresholds), and BOLD responses in each node of the functional network revealed in Chapter 6. Relevant BOLD responses for confidence judgements are (i) the difference between guess and confident responses, (not shown – all *n.s.*) and (ii) the effect of attention (shown). The former reflects sensitivity to subjective judgement whereas the latter reflects sensitivity to uncertainty or task-relevance (see Chapter 6). Results show that perceived self-efficacy is associated with attentional responses in left frontal pole, and marginally in right OFC.

Confidence is shaped by reward and value (De Martino et al., 2013; Hebart et al., 2014), which themselves are contextual and thus shaped by the perceptual systems. Thus, one possibility is that priors for confidence arise via interactions with decision-making mechanisms in seemingly parallel domains.

Another possibility recruits counterfactual predictions and sensorimotor contingencies (Seth, 2014a) into an ‘error-detection’ mechanism. To illustrate, suppose I see a figure in the fog. If I believe that figure to be cause by a fox running towards me, I may have a counterfactual prediction that at time $t + 1$, the fox will have advanced towards me at a fox-like speed. My posterior belief inferred at time $t + 1$ can then be compared to the posterior belief I would have expected, had the figure indeed been a fox. My prior in my decision accuracy can then be updated with the outcome of this comparison process. More formally, this process can be described as one in which counterfactual predictions associated with the posterior belief are tested against the world, and if those counterfactual predictions hold, the aforementioned posterior belief is likely to have been correct.

7.4 Future Directions

The empirical work in this thesis has revealed that confidence is strongly shaped by the extent to which perceptual decisions are supported by prior expectations. Perceptual expectations begin to shape subjective confidence prior to the appearance of a stimulus, and are integrated into confidence judgements in rIFG by comparing the associated perceptual decision against the prior evidence in its favour. Section 7.3 has proposed a Bayesian brain model, in which confidence is constructed from a ‘second-order’ inference, and that posits a mechanism for the construction of confidence from decision evidence and expected ‘self-efficacy’. However, many questions pertaining to the role of top-down influences in construction of confidence remain. This section will outline some directions for future research, namely, on the role of attention in confidence judgements, on the difference between confidence and uncertainty, and on whether non-perceptual priors shape confidence.

First, while the primary aim of manipulating attention here was to isolate effects of expectation, understanding the construction of confidence requires understanding the role of attention as well. The work here did not find a strong effect of attention on the influence of expectations. However, we know that sensory uncertainty and perceptual sensitivity have powerful effects on subjective visibility. Moreover, we know that even though the ideal Bayesian observer will use expectations more when sensory signals are imprecise (under inattention), empirical work finds that attention amplifies or optimises expectancy effects (see section 2.2.3), if it has any effect at all. So, why did attention have so few effects on subjective perception here?

One possibility is that the dual-task paradigm suppressed attentional effects that may otherwise have been present, because though the expected task demands were different, task difficulty was equated across trials. It may be that allowing top-down attention to shape the difficulty of the task leads to this perceived difficulty shaping decision confidence even after accounting for changes in perceptual sensitivity. Another possibility is that top-down attention is involved in inferring sensory noise, which was kept constant across participants in these studies. Recent work has shown that, as expected, confidence decreases with increasing sensory variance after equating sensitivity (Spence et al., 2015), yet variance is systematically underestimated (Zylberberg et al., 2014). This underestimation may be associated with changes in the reliability of prior evidence, as would be predicted by ideal Bayesian observer models. One avenue for future research could be to factorially manipulate the expected mean and variance of an RDK to determine how confidence is shaped by each of these predictive pieces of information. Investigating the role of top-down attention within such a factorial design may help elucidate its role in the construction of confidence.

It is clear that uncertainty – about the sensory signals, internal state or action outcomes – strongly shapes perceptual decision-making and confidence, and that these distinct forms of uncertainty are represented in process-specific brain regions (Bach & Dolan, 2012). The incorporation of uncertainty also tends to be optimally incorporated into objective decisions (Knill & Pouget, 2004). The studies in this thesis show that confidence in perceptual decision-making under sensory uncertainty is shaped by prior expectations, such that in the post-decision period confidence is represented as perceptual ‘prediction error’ in right inferior frontal gyrus. These results are explained by appealing to the notion that template-response matching contributes to the construction of confidence. However, on an alternative view expectations shape decision confidence only indirectly, with this effect being driven by changes in representational uncertainty (e.g. the variance of the posterior *pdf*). In other words, do expectations shape decision confidence over and above their effects on uncertainty, or can expectancy effects on decisional certainty account for all effects on decision confidence? Teasing apart choice confidence and decisional uncertainty may then reveal neural mechanisms that are implicated specifically in the construction of confidence, but not its antecedents or ensuing processes.

Finally, future research could investigate the role of prior beliefs that are not perceptual in nature. The model presented in section 7.6 proposes a critical role for priors about perceptual performance, or ‘self-efficacy’, yet the empirical work in this thesis has not manipulated these priors explicitly. Recent work has shown that optimism influences how priors are used in decision-making and updating beliefs: optimistic individuals are more prone to update beliefs on the basis of positive information (Sharot et al., 2011) and to have higher priors on reward (Stankevicius, Huys, Kalra, & Seriès, 2014). Similarly, encoding fluency – believing that a stimulus is easily learned - is associated with higher judgements of learning (the easily learned = easily remembered effect, Koriat 2008). This suggests that task-specific beliefs about performance are associated with higher confidence, at least in the memory domain. This effect is further constrained by the finding that an even more abstracted beliefs: believing that effortful decisions are a result of task ability is associated with a stronger relationship between encoding fluency and judgements of learning (Miele, Finn, & Molden, 2011).

So, would priors on perceptual performance shape perceptual confidence? In the memory domain, prospective confidence judgements do correlate with retrospective trial-by-trial confidence judgements, even in non-human primates (G. Morgan, Kornell, Kornblum, & Terrace, 2014), indicating that we can model prior beliefs about performance. Moreover, recent work has used reinforcement learning models to capture the effects of expected confidence – a function of recent confidence judgements – on trial-by-trial confidence (Guggenmos et al., 2016), revealing a key role of striatum in the representation of what is referred to here as expected 'self-efficacy'. Are prospective and retrospective perceptual confidence judgements represented in different brain areas? Can retrospective confidence be modelled as a function of perceptual prediction error and prospective confidence? Model-based fMRI could test these questions explicitly, by formulating retrospective confidence as a 'second-order' posterior: the probability of being correct, given the decision evidence ('first order' posterior) and prior (prospective confidence).

7.5 Conclusions

This thesis has addressed the question of whether and how perceptual prior expectations shape confidence judgements. Results show that subjective confidence increases with increasing prior probability of the decision. The process by which confidence is shaped by perceptual priors begins before stimulus onset, where the weighting of priors on decision and confidence is determined according to the phase of ongoing occipital alpha oscillations. Right inferior frontal gyrus then integrates neural responses to expectation-report mismatch into confidence signals, recruiting both visual and frontal regions. Together, these results show that top-down influences of expectation shape our perceptual experience in a similar manner to that seen in objective perception, such that we largely see what we believe to be true.

BIBLIOGRAPHY

Aitchison, L., Bang, D., Bahrami, B., & Latham, P. E. (2015). Doubly Bayesian Analysis of Confidence in Perceptual Decision-Making. *PLOS Computational Biology*, *11*, e1004519.

Allen, M., Fardo, F., Dietz, M. J., Hillebrandt, H., Friston, K. J., Rees, G., & Roepstorff, A. (2016). Anterior insula coordinates hierarchical processing of tactile mismatch responses. *NeuroImage*, *127*, 34–43.

Arnal, L. H., & Giraud, A.-L. (2012). Cortical oscillations and sensory predictions. *Trends in Cognitive Sciences*, *16*, 390–8.

Arnal, L. H., Wyart, V., & Giraud, A.-L. (2011). Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. *Nature Neuroscience*, *14*, 797–801.

Ashburner, J., & Friston, K. J. (2000). Voxel-based morphometry--the methods. *NeuroImage*, *11*, 805–21.

Bach, D. R., & Dolan, R. J. (2012). Knowing how much you don't know: a neural organization of uncertainty estimates. *Nature Reviews Neuroscience*, *13*, 572–586.

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science (New York, N.Y.)*, *329*, 1081–5.

Baird, B., Smallwood, J., Gorgolewski, K. J., & Margulies, D. S. (2013). Medial and lateral networks in anterior prefrontal cortex support metacognitive ability for memory and perception. *The Journal of Neuroscience*, *33*, 16657–65.

Banjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*, 289–300.

Bar, M. (2003). A Cortical Mechanism for Triggering Top-Down Facilitation in Visual Object Recognition. *Journal of Cognitive Neuroscience*, *15*, 600–609.

Bar, M. (2007). The proactive brain: using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, *11*, 280–289.

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., … Halgren, E. (2006). Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences*, *103*, 449–454.

Baranski, J. V, & Petrusic, W. M. (1998). Probing the locus of confidence judgments: experiments on the time to determine confidence. *Journal of Experimental Psychology. Human Perception and Performance*, *24*, 929–945.

Barrett, A. B., Dienes, Z., & Seth, A. K. (2013). Measures of Metacognition on Signal-Detection Theoretic Models. *Psychological Methods*. doi:10.1037/a0033268

Bastos, A. M., Usrey, W. M., Adams, R. a, Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, *76*, 695–711.

Bastos, A. M., Vezoli, J., Bosman, C. A., Schoffelen, J. M., Oostenveld, R., Dowdall, J. R., … Fries, P. (2015). Visual areas exert feedforward and feedback influences through distinct frequency channels. *Neuron*, *85*, 390–401.

Bauer, M., Stenner, M.-P., Friston, K. J., & Dolan, R. J. (2014). Attentional Modulation of Alpha/Beta and Gamma Oscillations Reflect Functionally Distinct Processes. *The Journal of Neuroscience*, *34*, 16117–16125.

Bayer, H. M., & Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, *47*, 129–141.

Beck, D. M., & Kastner, S. (2009). Top-down and bottom-up mechanisms in biasing competition in the human brain. *Vision Research*, *49*, 1154–65.

Beck, D. M., Rees, G., Frith, C. D., & Lavie, N. (2001). Neural correlates of change detection and change blindness. *Nature Neuroscience*, *4*, 645–650.

Beck, J. M., Ma, W. J., Kiani, R., & Hanks, T. (2008). Probabilistic population codes for Bayesian decision making. *Neuron*, *60*, 1142–1152.

Berens, P. (2009). CircStat: A MATLAB toolbox for circular statistics. *Journal of Statistical Software*, *31*.

Bitzer, S., Park, H., Blankenburg, F., & Kiebel, S. J. (2014). Perceptual decision making: drift-diffusion model is equivalent to a Bayesian model. *Frontiers in Human Neuroscience*, *8*, 102.

Blackwood, N., Ffytche, D., Simmons, A., Bentall, R., Murray, R., & Howard, R. (2004). The cerebellum and decision making under uncertainty. *Brain Research. Cognitive Brain Research*, *20*, 46–53.

Blakemore, S. J., Frith, C. D., & Wolpert, D. M. (2001). The cerebellum is involved in predicting the sensory consequences of action. *Neuroreport*, *12*, 1879–1884.

Boldt, A., & Yeung, N. (2015). Shared Neural Markers of Decision Confidence and Error Detection. *The Journal of Neuroscience*, *35*, 3478–3484.

Börgers, C., & Kopell, N. J. (2008). Gamma oscillations and stimulus selection. *Neural Computation*, *20*, 383–414.

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, *138*, 389–414.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436.

Braunlich, K., & Seger, C. A. (2016). Categorical evidence, confidence, and urgency during probabilistic categorization. *NeuroImage*, *125*, 941–952.

Brayanov, J. B., & Smith, M. a. (2010). Bayesian and “anti-Bayesian” biases in sensory integration for action and perception in the size-weight illusion. *Journal of Neurophysiology*, *103*, 1518–31.

Brodski, A., Paasch, G.-F., Helbling, S., & Wibral, M. (2015). The Faces of Predictive Coding. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *35*, 8997–9006.

Brown, H. R., & Friston, K. J. (2012). Dynamic causal modelling of precision and synaptic gain in visual perception - an EEG study. *NeuroImage*, *63*, 223–231.

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*, 153–178.

Brunet, N., Bosman, C. a, Vinck, M., Roberts, M., Oostenveld, R., Desimone, R., … Fries, P. (2014). Stimulus repetition modulates gamma-band synchronization in primate visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, 3626–31.

Bubic, A., von Cramon, D. Y., Jacobsen, T., Schröger, E., & Schubotz, R. I. (2009). Violation of expectation: neural correlates reflect bases of prediction. *Journal of*

Cognitive Neuroscience, *21*, 155–68.

Bülthoff, I., Bülthoff, H., & Sinha, P. (1998). Top-down influences on stereoscopic depth-perception. *Nature Neuroscience*, *1*, 254–257.

Busch, N. a, Dubois, J., & VanRullen, R. (2009). The phase of ongoing EEG oscillations predicts visual perception. *The Journal of Neuroscience*, *29*, 7869–76.

Busch, N. a, & VanRullen, R. (2010). Spontaneous EEG oscillations reveal periodic sampling of visual attention. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 16048–53.

Buzsáki, G., & Wang, X.-J. (2012). Mechanisms of gamma oscillations. *Annual Review of Neuroscience*, *35*, 203–25.

Casey, B. J., Thomas, K. M., Welsh, T. F., Badgaiyan, R. D., Eccard, C. H., Jennings, J. R., & Crone, E. A. (2000). Dissociation of response conflict , attentional selection , and expectancy with functional magnetic resonance imaging. *Proceedings of the National Academy of Sciences*, *97*, 8728–8733.

Chang, A. Y.-C., Kanai, R., & Seth, A. K. (2015). Cross-modal prediction changes the timing of conscious access during the motion-induced blindness. *Consciousness and Cognition*, *31*, 139–147.

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

Cheadle, S., Egner, T., Wyart, V., Wu, C., & Summerfield, C. (2015). Feature expectation heightens visual sensitivity during fine orientation discrimination. *Journal of Vision*, *15*, 14.

Chennu, S., Noreika, V., Gueorguiev, D., Blenkmann, A., Kochen, S., Ibáñez, A., … Bekinschtein, T. (2013). Expectation and attention in hierarchical auditory prediction. *The Journal of Neuroscience*, *33*, 11194–205.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *The Behavioral and Brain Sciences*, *36*, 181–204.

Clark, A. (2015). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind.* Oxford University Press.

Clarke, F. R., Birdsall, T. G., & Tanner, W. P. (1959). Two types of ROC curves and definitions of parameters. *The Journal of the Acoustical Society of America*, *31*, 629–630.

Cleeremans, A. (2011). The Radical Plasticity Thesis: How the Brain Learns to be Conscious. *Frontiers in Psychology*, *2*, 1–12.

Clifford, C. W. G., Arabzadeh, E., & Harris, J. a. (2008). Getting technical about awareness. *Trends in Cognitive Sciences*, *12*, 54–8.

Clos, M., Langner, R., Meyer, M., Oechslin, M. S., Zilles, K., & Eickhoff, S. B. (2014). Effects of prior information on decoding degraded speech: An fMRI study. *Human Brain Mapping*, *35*, 61–74.

Cooper, P. S., Darriba, Á., Karayanidis, F., & Barceló, F. (2016). Contextually sensitive power changes across multiple frequency bands underpin cognitive control. *NeuroImage*, *132*, 499–511.

Corbetta, M., & Shulman, G. L. (2002). Control of Goal-Directed and Stimulus-Driven Attention in the Brain. *Nature Reviews Neuroscience*, *3*, 215–229.

Coste, C. P., Sadaghiani, S., Friston, K. J., & Kleinschmidt, A. (2011). Ongoing brain activity fluctuations directly account for intertrial and indirectly for intersubject variability in Stroop task performance. *Cerebral Cortex (New York, N.Y.: 1991)*, *21*, 2612–9.

Criaud, M., & Boulinguez, P. (2013). Have we been asking the right questions when assessing response inhibition in go/no-go tasks with fMRI? A meta-analysis and critical review. *Neuroscience and Biobehavioral Reviews*, *37*, 11–23.

d'Acremont, M., Schultz, W., & Bossaerts, P. (2013). The Human Brain Encodes Event Frequencies While Forming Subjective Beliefs. *The Journal of Neuroscience*, *33*, 10887–10897.

Daniel, R., & Pollmann, S. (2012). Striatal activations signal prediction errors on confidence in the absence of external feedback. *NeuroImage*, *59*, 3457–3467.

Daunizeau, J., den Ouden, H. E. M., Pessiglione, M., Kiebel, S. J., Stephan, K. E., & Friston, K. J. (2010). Observing the observer (I): meta-bayesian models of learning and decision-making. *PloS One*, *5*, e15554.

de Gardelle, V., Le Corre, F., & Mamassian, P. (2016). Confidence as a Common Currency between Vision and Audition. *Plos One*, *11*, e0147901.

de Gardelle, V., & Mamassian, P. (2014). Does Confidence Use a Common Currency Across Two Visual Tasks? *Psychological Science*, *25*, 1286–1288.

de Gardelle, V., & Summerfield, C. (2011). Robust averaging during perceptual judgment. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 13341–6.

de Lange, F. P., Rahnev, D. A., Donner, T. H., & Lau, H. (2013). Prestimulus oscillatory activity over motor cortex reflects perceptual expectations. *The Journal of Neuroscience*, *33*, 1400–10.

De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nature Neuroscience*, *16*, 105–10.

den Ouden, H. E. M., Friston, K. J., Daw, N. D., McIntosh, a. R., & Stephan, K. E. (2009). A Dual Role for Prediction Error in Associative Learning. *Cerebral Cortex*, *19*, 1175–1185.

den Ouden, H. E. M., Kok, P., & de Lange, F. P. (2012). How Prediction Errors Shape Perception, Attention, and Motivation. *Frontiers in Psychology*, *3*, 1–12.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*, 193–222.

Dienes, Z. (2008). Subjective measures of unconscious knowledge. *Progress in Brain Research*, *168*, 49–64.

Dienes, Z., & Seth, A. K. (2010). Measuring any conscious content versus measuring the relevant conscious content: Comment on Sandberg et al . q. *Consciousness and Cognition*, *19*, 1079–1080.

Domenech, P., & Dreher, J.-C. (2010). Decision threshold modulation in the human brain. *The Journal of Neuroscience*, *30*, 14305–14317.

Donner, T. H., & Siegel, M. (2011). A framework for local cortical oscillation patterns. *Trends in Cognitive Sciences*, *15*, 191–9.

Downar, J., Crawley, a P., Mikulis, D. J., & Davis, K. D. (2000). A multimodal cortical network for the detection of changes in the sensory environment. *Nature Neuroscience*, *3*, 277–283.

Downar, J., Crawley, A. P., Mikulis, D. J., & Davis, K. D. (2002). A cortical network sensitive to stimulus salience in a neutral behavioral context across multiple sensory modalities. *Journal of Neurophysiology*, *87*, 615–620.

Drewes, J., & VanRullen, R. (2011). This is the rhythm of your eyes: the phase of ongoing electroencephalogram oscillations modulates saccadic reaction time. *The Journal of Neuroscience*, *31*, 4698–708.

Dugué, L., Marque, P., & VanRullen, R. (2011). The phase of ongoing oscillations mediates the causal relation between brain excitation and visual perception. *The Journal of Neuroscience*, *31*, 11889–93.

Duncan, J. (2006). EPS Mid-Career Award 2004: brain mechanisms of attention. *The Quarterly Journal of Experimental Psychology (2006)*, *59*, 2–27.

Dunovan, K. E., Tremel, J. J., & Wheeler, M. E. (2014). Prior probability and feature predictability interactively bias perceptual decisions. *Neuropsychologia*, *61*, 210–221.

Ebner, T. J., & Pasalar, S. (2008). Cerebellum predicts the future motor state. *Cerebellum*, *7*, 583–588.

Edwards, M. J., Adams, R. A., Brown, H., Pareés, I., & Friston, K. J. (2012). A Bayesian account of “hysteria.” *Brain*, *135*, 3495–3512.

Egner, T., Monti, J. M., & Summerfield, C. (2010). Expectation and surprise determine neural population responses in the ventral visual stream. *The Journal of Neuroscience*, *30*, 16601–16608.

Eickhoff, S. B., Pomjanski, W., Jakobs, O., Zilles, K., & Langner, R. (2011). Neural correlates of developing and adapting behavioral biases in speeded choice reactions-An fMRI study on predictive motor coding. *Cerebral Cortex*, *21*, 1178–1191.

Eickhoff, S. B., Stephan, K. E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., & Zilles, K. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *Neuroimage*, *25*, 1325–1335.

Engel, A. K., & Fries, P. (2010). Beta-band oscillations--signalling the status quo? *Current Opinion in Neurobiology*, *20*, 156–65.

Engel, A. K., Fries, P., & Singer, W. (2001). Dynamic predictions: oscillations and synchrony in top-down processing. *Nature Reviews Neuroscience*, *2*, 704–716.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*, 429–433.

Evans, S., & Azzopardi, P. (2007). Evaluation of a “bias-free” measure of awareness. *Spatial Vision*, *20*, 61–77.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral and biomedical sciences. *Behavior Research Methods*, *39*, 175–191.

Feldman, H., & Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, *4*, 215.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, *116*, 752–782.

Fernandez-Duque, D., Baird, J., & Posner, M. (2000). Awareness and metacognition. *Consciousness and Cognition*, *9*, 324–6.

Festinger, L. (1943). Studies in decision: I. Decision-time, relative frequency of judgment and subjective confidence as related to physical stimulus difference.

Journal of Experimental Psychology, *32*, 291.

Fetsch, C. R., Kiani, R., Newsome, W. T., & Shadlen, M. N. (2014). Effects of Cortical Microstimulation on Confidence in a Perceptual Decision. *Neuron*, *83*, 797–804.

Fetsch, C. R., Kiani, R., & Shadlen, M. N. (2015). Predicting the Accuracy of a Decision: A Neural Mechanism of Confidence. *Cold Spring Harbor Symposia on Quantitative Biology*, *79*, 185–197.

Finley, J. R., Benjamin, A. S., & Mccarley, J. S. (2014). Metacognition of multi-tasking: how well do we predict the costs of divided attention?, *20*, 158–165.

Firestone, C., & Scholl, B. J. (2015). Cognition does not affect perception: Evaluating the evidence for “top-down” effects. *Behavioral and Brain Sciences*, *4629*, 1–77.

Fiser, J., Berkes, P., Orbán, G., & Lengyel, M. (2010). Statistically optimal perception and learning: from behavior to neural representations. *Trends in Cognitive Sciences*, *14*, 119–130.

Fleck, M. S., Daselaar, S. M., Dobbins, I. G., & Cabeza, R. (2006). Role of prefrontal and anterior cingulate regions in decision-making processes shared by memory and nonmemory tasks. *Cerebral Cortex (New York, N.Y. : 1991)*, *16*, 1623–30.

Fleming, S. M., & Dolan, R. J. (2012). Neural basis of metacognition. *Philosophical Transactions of the Royal Society B Biological Sciences*, *367*, 1338–1349.

Fleming, S. M., Huijgen, J., & Dolan, R. J. (2012). Prefrontal contributions to metacognition in perceptual decision making. *The Journal of Neuroscience*, *32*, 6117–25.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, 1–9.

Fleming, S. M., Ryu, J., Golfinos, J. G., & Blackmon, K. E. (2014). Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain : A Journal of Neurology*, *137*, 2811–2822.

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science (New York, N.Y.)*, *329*, 1541–1543.

Forstmann, B. U., Ratcliff, R., & Wagenmakers, E.-J. (2016). Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. *Annual Review of Psychology*, *67*, 1–26.

Frey, J. N., Ruhnau, P., & Weisz, N. (2015). Not so different after all: The same oscillatory processes support different types of attention. *Brain Research*, 1–14.

Fries, P. (2005). A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends in Cognitive Sciences*, *9*, 474–80.

Friston, K. (2014). Active inference and agency. *Cognitive Neuroscience*, *5*, 119–21.

Friston, K. J. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, *13*, 293–301.

Friston, K. J. (2010). The free-energy principle: a unified brain theory? *Nature Reviews. Neuroscience*, *11*, 127–38.

Friston, K. J. (2012a). Predictive coding, precision and synchrony. *Cognitive Neuroscience*, *3*, 238–9.

Friston, K. J. (2012b). The history of the future of the Bayesian brain. *NeuroImage*, *62*, 1230–3.

Friston, K. J., Adams, R., Perrinet, L., & Breakspear, M. (2012). Perceptions as hypotheses: saccades as experiments. *Frontiers in Psychology*, *3*, 151.

Friston, K. J., & Ao, P. (2012). Free energy, value, and attractors. *Computational and Mathematical Methods in Medicine*, *2012*, 937860.

Friston, K. J., Bastos, A. M., Pinotsis, D., & Litvak, V. (2014). LFP and oscillations-what do they tell us? *Current Opinion in Neurobiology*, *31C*, 1–6.

Friston, K. J., & Kiebel, S. (2009a). Cortical circuits for perceptual inference. *Neural Networks : The Official Journal of the International Neural Network Society*, *22*, 1093–104.

Friston, K. J., & Kiebel, S. (2009b). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *364*, 1211–21.

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., FitzGerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience*, 150217111908007.

Galvin, S. J., Podd, J. V, Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review*, *10*, 843–76.

Garrido, M. I., Kilner, J. M., Kiebel, S. J., & Friston, K. J. (2009). Dynamic causal modeling of the response to frequency deviants. *Journal of Neurophysiology*, *101*, 2620–2631.

Gennaro, R. J. (1996). *Consciousness and self-consciousness: A defense of the higher-order thought theory of consciousness*. John Benjamins Publishing.

Gennaro, R. J. (2004). Higher-Order Theories of Consciousness : An Overview, 1–15.

Gherman, S., & Philiastides, M. G. (2015). Neural representations of confidence emerge from the process of decision formation during perceptual choices. *NeuroImage*, *106*, 134–143.

Gilbert, C. D., & Li, W. (2013). Top-down influences on visual processing. *Nature Reviews Neuroscience*, *14*, 350–363.

Gilbert, C. D., & Sigman, M. (2007). Brain states: top-down influences in sensory processing. *Neuron*, *54*, 677–96.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the Americal Statistical Association*, *49*, 732–764.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1). New York: Wiley.

Green, N., Bogacz, R., Huebl, J., Beyer, A. K., Kühn, A. A., & Heekeren, H. R. (2013). Reduction of influence of task difficulty on perceptual decision making by stn deep brain stimulation. *Current Biology*, *23*, 1681–1684.

Gregory, R. L. (1980). Perceptions as Hypotheses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *290*, 181–197.

Grimaldi, P., Lau, H., & Basso, M. A. (2015). There are things that we know that we know, and there are things that we do not know we do not know: Confidence in decision-making. *Neuroscience and Biobehavioral Reviews*, *55*, 88–97.

Grinband, J., Hirsch, J., & Ferrera, V. P. (2006). A neural representation of categorization uncertainty in the human brain. *Neuron*, *49*, 757–763.

Guggenmos, M., Wilbertz, G., Hebart, M. N., & Sterzer, P. (2016). Mesolimbic confidence signals guide perceptual learning in the absence of external feedback. *eLife*, *5*, 1–19.

Hampshire, A., Chamberlain, S. R., Monti, M. M., Duncan, J., & Owen, A. M. (2010). The role of the right inferior frontal gyrus: inhibition and attentional control. *NeuroImage*, *50*, 1313–9.

Hangya, B., Sanders, J. I., & Kepecs, A. (2016). A mathematical framework for statistical decision confidence. *bioRxiv*. Retrieved from http://biorxiv.org/content/early/2016/01/01/017400.abstract

Hansen, K. a., Hillenbrand, S. F., & Ungerleider, L. G. (2012). Effects of prior knowledge on decisions made under perceptual vs. Categorical uncertainty. *Frontiers in Neuroscience*, *6*, 1–10.

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values ofd′. *Behavior Research Methods, Instruments, & Computers*, *27*, 46–51.

Hebart, M. N., Schriever, Y., Donner, T. H., & Haynes, J.-D. (2014). The Relationship between Perceptual Decision Variables and Confidence in the Human Brain. *Cerebral Cortex (New York, N.Y. : 1991)*, bhu181–.

Helmholtz, H. V. (1860). Theorie der Luftschwingungen in Röhren mit offenen Enden. *Journal Für Die Reine Und Angewandte Mathematik*, 1–72.

Hesselmann, G., Kell, C. a, & Kleinschmidt, A. (2008). Ongoing activity fluctuations in hMT+ bias the perception of coherent visual motion. *The Journal of Neuroscience*, *28*, 14481–5.

Hesselmann, G., Sadaghiani, S., Friston, K. J., & Kleinschmidt, A. (2010). Predictive coding or evidence accumulation? False inference and neuronal fluctuations. *PloS One*, *5*, e9926.

Hickey, C., & Theeuwes, J. (2011). Context and competition in the capture of visual attention. *Attention, Perception & Psychophysics*, *73*, 2053–64.

Hilgenstock, R., Weiss, T., & Witte, O. W. (2014). You'd Better Think Twice: Post-Decision Perceptual Confidence. *NeuroImage*, *99*, 323–331.

Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology*, *3*, 96.

Hohwy, J. (2013). *The Predictive Mind*. Oxford: OUP.

Hohwy, J., Roepstorff, A., & Friston, K. J. (2008). Predictive coding explains binocular rivalry: an epistemological review. *Cognition*, *108*, 687–701.

Horga, G., Schatz, K. C., Abi-Dargham, A., & Peterson, B. S. (2014). Deficits in predictive coding underlie hallucinations in schizophrenia. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *34*, 8072–82.

Hoyer, P. O., & Hyvarinen, A. (2003). Interpreting neural response variability as Monte Carlo sampling of the posterior. *Advances in Neural Information Processing Systems*, 293–300.

Hsu, Y.-F., Hämäläinen, J. a, & Waszak, F. (2014a). Both attention and prediction are necessary for adaptive neuronal tuning in sensory processing. *Frontiers in Human Neuroscience*, *8*, 152.

Hsu, Y.-F., Hämäläinen, J. a, & Waszak, F. (2014b). Repetition suppression comprises both attention-independent and attention-dependent processes. *NeuroImage*, *98*, 168–75.

Iglesias, S., Mathys, C., Brodersen, K. H., Kasper, L., Piccirelli, M., den Ouden, H. E., & Stephan, K. E. (2013). Hierarchical Prediction Errors in Midbrain and Basal Forebrain during Sensory Learning. *Neuron*, *80*, 519–530.

Jachs, B., Blanco, M. J., Grantham-Hill, S., & Soto, D. (2015). On the Independence of Visual Awareness and Metacognition : A Signal Detection Theoretic Analysis. *Journal of Experimental Psychology : Human Perception and Performance*, *41*, 269–276.

Jaramillo, S., & Zador, A. M. (2011). The auditory cortex mediates the perceptual effects of acoustic temporal expectation. *Nature Neuroscience*, *14*, 246–51.

Jensen, O., Bonnefond, M., & VanRullen, R. (2012). An oscillatory mechanism for prioritizing salient unattended stimuli. *Trends in Cognitive Sciences*, *16*, 200–6.

Jiang, J., Summerfield, C., & Egner, T. (2013). Attention sharpens the distinction between expected and unexpected percepts in the visual brain. *The Journal of Neuroscience*, *33*, 18438–47.

Kanai, R., Komura, Y., Shipp, S., & Friston, K. (2015). Cerebral hierarchies : predictive processing , precision and the pulvinar. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *370*.

Kanai, R., Muggleton, N. G., & Walsh, V. (2008). TMS over the intraparietal sulcus induces perceptual fading. *Journal of Neurophysiology*, *100*, 3343–50.

Kanai, R., Tsuchiya, N., & Verstraten, F. a J. (2006). The scope and limits of top-down attention in unconscious visual processing. *Current Biology : CB*, *16*, 2332–6.

Kanai, R., Walsh, V., & Tseng, C. (2010). Subjective discriminability of invisibility: a framework for distinguishing perceptual and attentional failures of awareness. *Consciousness and Cognition*, *19*, 1045–1057.

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the Americal Statistical Association*, *90*, 773–795.

Kentridge, R. W., & Heywood, C. a. (2000). Metacognition and awareness. *Consciousness and Cognition*, *9*, 308–12; discussion 324–6.

Kepecs, A., & Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*, 1322–37.

Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature*, *455*, 227–31.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object Perception as Bayesian Inference. *Annual Review of Psychology*, *55*, 271–304.

Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice Certainty Is Informed by Both Evidence and Decision Time. *Neuron*, *84*, 1329–1342.

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

Kim, H., Shimojo, S., & O’Doherty, J. P. (2011). Overlapping Responses for the Expectation of Juice and Money Rewards in Human Ventromedial Prefrontal Cortex. *Cerebral Cortex*, *21*, 769–776.

Kleiner, M., D., B., & Pelli, D. (2007). What’s new in Psychtoolbox-3? In *Perception 36 ECVP Abstract Supplement*.

Klimesch, W., Sauseng, P., & Hanslmayr, S. (2007). EEG alpha oscillations: the inhibition-timing hypothesis. *Brain Research Reviews*, *53*, 63–88.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*, 712–9.

Knudsen, E. I. (2007). Fundamental components of attention. *Annual Review of Neuroscience*, *30*, 57–78.

Ko, Y., & Lau, H. (2012). A detection theoretic explanation of blindsight suggests a link between conscious perception and metacognition. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*, 1401–11.

Koerding, K. P., Ku, S.-P., & Wolpert, D. M. (2004). Bayesian estimation in force integration. *Journal of Neurophysiology*, *92*, 3161–3165.

Kok, P., Brouwer, G. J., van Gerven, M. a. J., & de Lange, F. P. (2013). Prior Expectations Bias Sensory Representations in Visual Cortex. *The Journal of Neuroscience*, *33*, 16275–16284.

Kok, P., Jehee, J. F., & de Lange, F. P. (2012). Less Is More : Expectation Sharpens Representations in the Primary Visual Cortex. *Neuron*, *75*, 265–270.

Kok, P., Rahnev, D., Jehee, J. F. M., Lau, H. C., & de Lange, F. P. (2011). Attention reverses the effect of prediction in silencing sensory signals. *Cerebral Cortex (New York, N.Y. : 1991)*, *22*, 2197–206.

Kopell, N., Kramer, M. a, Malerba, P., & Whittington, M. a. (2010). Are different rhythms good for different functions? *Frontiers in Human Neuroscience*, *4*, 187.

Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*, 244–247.

Koriat, A. (2008). Easy comes, easy goes? The link between learning and remembering and its exploitation in metacognition. *Memory & Cognition*, *36*, 416–428.

Koster-Hale, J., & Saxe, R. (2013). Theory of Mind: A Neural Prediction Problem. *Neuron*.

Kouider, S., Long, B., Le Stanc, L., Charron, S., Fievet, A.-C., Barbosa, L. S., & Gelskov, S. V. (2015). Neural dynamics of prediction and surprise in infants. *Nat Commun*, *6*, 8537.

Kunimoto, C., Miller, J., & Pashler, H. (2001). Confidence and accuracy of near-threshold discrimination responses. *Consciousness and Cognition*, *10*, 294–340.

Kvam, P. D., Pleskac, T. J., Yu, S., & Busemeyer, J. R. (2015). Interference effects of choice on confidence: Quantum characteristics of evidence accumulation. *Proceedings of the National Academy of Sciences of the United States of America*, *112*, 10645–10650.

Landau, A. N., & Fries, P. (2012). Attention samples stimuli rhythmically. *Current Biology : CB*, *22*, 1000–4.

Larsson, J., & Smith, A. T. (2012). fMRI repetition suppression: neuronal adaptation or stimulus expectation? *Cerebral Cortex (New York, N.Y. : 1991)*, *22*, 567–76.

Lau, H. C. (2007). A higher order Bayesian decision theory of consciousness. *Prog.Brain Res.*, *168*, 35–48.

Lau, H. C., & Passingham, R. E. (2006). Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proceedings of the National Academy of Sciences of the United States of America*, *103*, 18763–18768.

Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences*, *15*, 365–73.

Lavie, N. (2006). The role of perceptual load in visual awareness. *Brain Research*, *1080*, 91–100.

Law, J. R., Flanery, M. a, Wirth, S., Yanike, M., Smith, A. C., Frank, L. M., … Stark, C. E. L. (2005). Functional magnetic resonance imaging activity during the gradual acquisition and expression of paired-associate memory. *The Journal of Neuroscience*, *25*, 5720–5729.

Lebreton, M., Abitbol, R., Daunizeau, J., & Pessiglione, M. (2015). Automatic integration of confidence in the brain valuation signal. *Nature Neuroscience*, *18*, 1159–1167.

Lee, T. S. (2002). Top-down influence in early visual processing: a Bayesian perspective. *Physiology & Behavior*, *77*, 645–650.

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A*, *20*, 1434.

Leopold, D. a. (2012). Primary visual cortex: awareness and blindsight. *Annual Review of Neuroscience*, *35*, 91–109.

Lindsley, D. B. (1952). Psychological phenomena and the electroencephalogram. *Electroencephalography and Clinical Neurophysiology*, *4*, 443–456.

Locke, J. (1700). *An essay concerning human understanding*.

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, A. J., Ly, A., Gronau, Q. F., Smira, M., Epskamp, S., Matzke, D., Wild, A., Rouder, J. N., Morey, R. D. & Wagenmakers, E.-J. (2015). JASP (Version 0.7).

Luck, S. J. (2005). *An introduction to the event-related potential technique*. Cambridge, M.A.: MIT Press.

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*, 1432–8.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Psychology Press.

Macmillan, N. A., & Kaplan, H. L. (1985). Detection theory analysis of group data: Estimating sensitivity from average hit and false-alarm rates. *Psychological Bulletin*, *98*, 185–199.

Maloney, L. T., Dal Martello, M. F., Sahm, C., & Spillmann, L. (2005). Past trials influence perception of ambiguous motion quartets through pattern completion. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 3164–3169.

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*, 422–30.

Maniscalco, B., & Lau, H. (2014). Signal Detection Theory Analysis of Type 1 and Type 2 Data: Meta-d′, Response-Specific Meta-d′, and the Unequal Variance SDT Model. In *The Cognitive Neuroscience of Metacognition* (pp. 25–66). Berlin Heidelberg: Springer.

Maniscalco, B., Peters, M. A. K., & Lau, H. (2016). Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity. *Attention, Perception & Psychophysics*, *78*, 923–937.

Mansfield, E. L., Karayanidis, F., Jamadar, S., Heathcote, A., & Forstmann, B. U. (2011). Adjustments of response threshold during task switching: a model-based functional magnetic resonance imaging study. *The Journal of Neuroscience*, *31*, 14688–92.

Mansouri, F. a, Tanaka, K., & Buckley, M. J. (2009). Conflict-induced behavioural adjustment: a clue to the executive functions of the prefrontal cortex. *Nature Reviews. Neuroscience*, *10*, 141–52.

Marcel, A. J. (1993). Slippage in the unity of consciousness. In G. R. Bock & J. Marsh (Eds.), *Experimental and theoretical studies of consciousness.* (Vol. 174, pp. 166–168). John Whiley & Sons.

Martinez-Trujillo, J. C., & Treue, S. (2004). Feature-based attention increases the selectivity of population responses in primate visual cortex. *Current Biology*, *14*, 744–751.

Masson, M. E. J., & Rotello, C. M. (2009). Sources of bias in the Goodman-Kruskal gamma coefficient measure of association: implications for studies of metacognitive processes. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*, 509–27.

Mathewson, K. E., Gratton, G., Fabiani, M., Beck, D. M., & Ro, T. (2009). To see or not to see: prestimulus alpha phase predicts visual awareness. *The Journal of Neuroscience*, *29*, 2725–32.

Mathewson, K. E., Prudhomme, C., Fabiani, M., Beck, D. M., Lleras, A., & Gratton, G. (2012). Making waves in the stream of consciousness: entraining oscillations in EEG alpha and fluctuations in visual awareness with rhythmic visual stimulation. *Journal of Cognitive Neuroscience*, *24*, 2321–33.

Mathys, C. D., Lomakina, E. I., Daunizeau, J., Iglesias, S., Brodersen, K. H., Friston, K. J., & Stephan, K. E. (2014). Uncertainty in perception and the Hierarchical Gaussian Filter. *Frontiers in Human Neuroscience*, *8*, 1–24.

Mayer, A., Schwiedrzik, C. M., Wibral, M., Singer, W., & Melloni, L. (2015). Expecting to See a Letter : Alpha Oscillations as Carriers of Top-Down Sensory Predictions. *Cerebral Cortex*, 1–15.

Mazaheri, A., DiQuattro, N. E., Bengson, J., & Geng, J. J. (2011). Pre-stimulus activity predicts the winner of top-down vs. bottom-up attentional selection. *PloS One*, *6*, e16243.

McAdams, C. J., & Maunsell, J. H. (2000). Attention to both space and feature modulates neuronal responses in macaque area V4. *Journal of Neurophysiology*, *83*, 1751–5.

McCurdy, L. Y., Maniscalco, B., Metcalfe, J., Liu, K. Y., de Lange, F. P., & Lau, H. (2013). Anatomical coupling between distinct metacognitive systems for memory and visual perception. *The Journal of Neuroscience*, *33*, 1897–906.

Mealor, A. D., & Dienes, Z. (2013). The speed of metacognition: taking time to get to know one's structural knowledge. *Consciousness and Cognition*, *22*, 123–36.

Melloni, L., Schwiedrzik, C. M., Müller, N., Rodriguez, E., & Singer, W. (2011). Expectations change the signatures and timing of electrophysiological correlates of perceptual awareness. *The Journal of Neuroscience*, *31*, 1386–96.

Meyniel, F., Schlunegger, D., & Dehaene, S. (2015). The Sense of Confidence during Probabilistic Learning: A Normative Account. *PLoS Computational Biology*, *11*, e1004305.

Meyniel, F., Sigman, M., & Mainen, Z. F. (2015). Confidence as Bayesian Probability: From Neural Origins to Behavior. *Neuron*, *88*, 78–92.

Miele, D. B., Finn, B., & Molden, D. C. (2011). Does easily learned mean easily remembered?: it depends on your beliefs about intelligence. *Psychological Science : A Journal of the American Psychological Society / APS*, *22*, 320–324.

Morales, J., Solovey, G., Maniscalco, B., Rahnev, D., de Lange, F. P., & Lau, H. (2015). Low attention impairs optimal incorporation of prior knowledge in perceptual decisions. *Attention, Perception & Psychophysics*, *77*, 2021–36.

Morgan, A. T., Petro, L. S., & Muckli, L. (2016). Cortical feedback to V1 and V2 contains unique information about high-level scene structure. *BioRxiv*.

Morgan, G., Kornell, N., Kornblum, T., & Terrace, H. S. (2014). Retrospective and Prospective Metacognitive Judgments in Rhesus Macaques (Macaca mulatta) Gin. *Animal Cognition*, *17*, 249–257.

Morris, R. W., Dezfouli, A., Griffiths, K. R., & Balleine, B. W. (2014). Action-value comparisons in the dorsolateral prefrontal cortex control choice between goal-directed actions. *Nature Communications*, *5*, 4390.

Muckli, L., De Martino, F., Vizioli, L., Petro, L. S., Smith, F. W., Ugurbil, K., … Yacoub, E. (2015). Contextual Feedback to Superficial Layers of V1. *Current Biology*, *25*, 2690–2695.

Mulder, M. J., van Maanen, L., & Forstmann, B. U. (2014). Perceptual decision neurosciences - A model-based review. *Neuroscience*, *277*, 872–884.

Murphy, P. R., Robertson, I. H., Harty, S., & O'Connell, R. G. (2015). Neural evidence accumulation persists after choice to inform metacognitive judgments. *eLife*. doi:10.7554/eLife.11946

Nakahara, H., Itoh, H., Kawagoe, R., Takikawa, Y., & Hikosaka, O. (2004). Dopamine Neurons Can Represent Context-Dependent Prediction Error. *Neuron*, *41*, 269–280.

Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An Approximately Bayesian Delta-Rule Model Explains the Dynamics of Belief Updating in a Changing Environment. *The Journal of Neuroscience*, *30*, 12366–12378.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, *95*, 109–33.

Okun, M., Steinmetz, N. A., Cossell, L., Iacaruso, M. F., Ko, H., Barthó, P., … Harris, K. D. (2015). Diverse coupling of neurons to populations in sensory cortex. *Nature*, *521*, 511–515.

Overgaard, M., & Sandberg, K. (2012). Kinds of access: different methods for report reveal different kinds of metacognitive access. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*, 1287–96.

Pajani, A., Kok, P., Kouider, S., & de Lange, F. P. (2015). Spontaneous Activity Patterns in Primary Visual Cortex Predispose to Visual Hallucinations. *The Journal of Neuroscience*, *35*, 12947–12953.

Palva, S., & Palva, J. M. (2007). New vistas for alpha-frequency band oscillations. *Trends in Neurosciences*, *30*, 150–8.

Paulin, M. G. (2005). Evolution of the cerebellum as a neuronal machine for Bayesian state estimation. *Journal of Neural Engineering*, *2*, S219–S234.

Peirce, C. S., & Jastrow, J. (1884). On small differences in sensation. *Memoirs of the*

National Academy of Sciences, *3*, 73–83.

Pelt, S. van, Heil, L., Kwisthout, J., Ondobaka, S., Rooij, I. van, & Bekkering, H. (2016). Beta- and gamma-band activity reflect predictive coding in the processing of causal events. *Social*, 1–8.

Perricone, P. (2011). *A taxonomy of metacognition*. Psychology Press.

Persaud, N., McLeod, P., & Cowey, A. (2007). Post-decision wagering objectively measures awareness. *Nature Neuroscience*, *10*, 257–61.

Petrides, M., & Pandya, D. N. (2002). Comparative cytoarchitectonic analysis of the human and the macaque ventrolateral prefrontal cortex and corticocortical connection patterns in the monkey. *European Journal of Neuroscience*, *16*, 291–310.

Petro, L. S., Vizioli, L., & Muckli, L. (2014). Contributions of cortical feedback to sensory processing in primary visual cortex. *Frontiers in Psychology*, *5*, 1–8.

Petrusic, W. M., & Baranski, J. V. (2003). Judging confidence influences decision processing in comparative judgments. *Psychonomic Bulletin & Review*, *10*, 177–183.

Pezzulo, G. (2012). An Active Inference view of cognitive control. *Frontiers in Psychology*, *3*. doi:10.3389/fpsyg.2012.00478

Pinto, Y., van Gaal, S., de Lange, F. P., Lamme, V. A. F., & Seth, A. K. (2015). Expectations accelerate entry of visual stimuli into awareness. *Journal of Vision*, *15*, 13.

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, *117*, 864–901.

Pollack, I. (1959). On Indices of Signal and Response Discriminability. *Journal of the Acoustical Society of America*, *31*, 1031.

Porter, P. (1954). Another picture-puzzle. *American Journal of Psychology*, *67*, 550–551.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, *32*, 3–25.

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty : distinct probabilistic quantities for different goals. *Nature Neuroscience*, *19*, 366–374.

Rahnev, D. a, Bahdo, L., de Lange, F. P., & Lau, H. (2012). Prestimulus hemodynamic activity in dorsal attention network is negatively associated with decision confidence in visual perception. *Journal of Neurophysiology*, *108*, 1529–36.

Rahnev, D., Koizumi, A., Mccurdy, L. Y., Esposito, M. D., & Lau, H. (2015). Confidence Leak in Perceptual Decision Making. *Psychological Science*, *26*, 1664–1680.

Rahnev, D., Maniscalco, B., Graves, T., Huang, E., de Lange, F. P., & Lau, H. (2011). Attention induces conservative subjective biases in visual perception. *Nature Neuroscience*, *14*, 1513–5.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*, 79–87.

Rao, R. P. N., & Ballard, D. H. (2004). Probabilistic Models of Attention based on Iconic Representations and Predictive Coding. In *Neurobiology of Attention*. New York: Academic Press.

Ratcliff, R., & McKoon, G. (2008). The Diffusion Decision Model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*, 873–922.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in Cognitive Sciences*, *20*, 260–281.

Ratcliff, R., & Starns, J. J. (2009). Modelling confidence and response time in recognition memory. *Psychological Review*, *116*, 59–83.

Ratcliff, R., & Starns, J. J. (2013). Modeling confidence judgments, response times, and multiple choices in decision making: recognition memory and motion discrimination. *Psychological Review*, *120*, 697–719.

Rauss, K., Schwartz, S., & Pourtois, G. (2011). Top-down effects on early visual processing in humans: a predictive coding framework. *Neuroscience and Biobehavioral Reviews*, *35*, 1237–53.

Rohenkohl, G., & Nobre, A. C. (2011). A Oscillations Related To Anticipatory Attention Follow Temporal Expectations. *The Journal of Neuroscience*, *31*, 14076–84.

Rolls, E. T., Grabenhorst, F., & Deco, G. (2010). Decision-making, errors, and confidence in the brain. *Journal of Neurophysiology*, *104*, 2359–74.

Rorden, C., & Brett, M. (2000). Stereotaxic display of brain lesions. *Behavioural Neurology*, *12*, 191–200.

Rosenthal, D. M. (1986). Two concepts of consciousness. *Philosophical Studies*, *49*, 329–359.

Rosenthal, D. M. (2000). Consciousness, content, and metacognitive judgments. *Consciousness and Cognition*, *9*, 203–14.

Roth, M. J., Synofzik, M., & Lindner, A. (2013). The cerebellum optimizes perceptual predictions about external sensory events. *Current Biology*, *23*, 930–935.

Rounis, E., Maniscalco, B., Rothwell, J. C., Passingham, R. E., & Lau, H. (2010). Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. *Cognitive Neuroscience*, *1*, 165–175.

Saenz, M., Buracas, G. T., & Boynton, G. M. (2002). Global effects of feature-based attention in human visual cortex. *Nature Neuroscience*, *5*, 631–632.

Sakagami, M., & Pan, X. (2007). Functional role of the ventrolateral prefrontal cortex in decision making. *Current Opinion in Neurobiology*, *17*, 228–233.

Samaha, J., Bauer, P., Cimaroli, S., & Postle, B. R. (2015). Top-down control of the phase of alpha-band oscillations as a mechanism for temporal prediction. *Proceedings of the National Academy of Sciences*, 201503686.

Sandberg, K., Timmermans, B., Overgaard, M., & Cleeremans, A. (2010). Measuring consciousness: is one measure better than the other? *Consciousness and Cognition*, *19*, 1069–78.

Sarter, M., Givens, B., & Bruno, J. P. (2001). The cognitive neuroscience of sustained attention: where top-down meets bottom-up. *Brain Research. Brain Research Reviews*, *35*, 146–60.

Schmack, K., Gòmez-Carrillo de Castro, A., Rothkirch, M., Sekutowicz, M., Rössler, H., Haynes, J.-D., … Sterzer, P. (2013). Delusions and the role of beliefs in perceptual inference. *The Journal of Neuroscience*, *33*, 13701–12.

Schoenbaum, G., & Roesch, M. (2005). Orbitofrontal cortex, associative learning, and expectancies. *Neuron*, *47*, 633–6.

Schultz, W., & Dickinson, A. (2000). Neuronal Coding. *Annual Review of Neuroscience*, *23*, 473–500.

Schwiedrzik, C. M., Singer, W., & Melloni, L. (2011). Subjective and objective learning effects dissociate in space and in time. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 4506–11.

Scott, R. B., Dienes, Z., Barrett, A. B., Bor, D., & Seth, A. K. (2014). Blind Insight: Metacognitive Discrimination Despite Chance Task Performance. *Psychological Science*, *25*, 2199–2208.

Sedley, W., Gander, P. E., Kumar, S., Kovach, C. K., Kawasaki, H., Iii, M. A. H., & Griffiths, T. D. (2016). Neural Signatures of Perceptual Inference. *eLife*, 1–13.

Serences, J. T., & Boynton, G. M. (2007). The representation of behavioral choice for motion in human visual cortex. *The Journal of Neuroscience*, *27*, 12893–9.

Seriès, P., & Seitz, A. R. (2013). Learning what to expect (in visual perception). *Frontiers in Human Neuroscience*, *7*, 1–14.

Seth, A. K. (2014a). A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia. *Cognitive Neuroscience*, *5*, 97–118.

Seth, A. K. (2014b). A predictive processing theory of sensorymotor contingencies: Explaining the puzzle of perceptual presence and absence in synaesthesia. *Cognitive Neuroscience*.

Seth, A. K., Dienes, Z., Cleeremans, A., Overgaard, M., & Pessoa, L. (2008). Measuring consciousness: relating behavioural and neurophysiological approaches. *Trends in Cognitive Sciences*, *12*, 314–21.

Seth, A. K., Suzuki, K., & Critchley, H. D. (2011). An interoceptive predictive coding model of conscious presence. *Frontiers in Psychology*, *2*, 395.

Sharot, T., Korn, C. W., & Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. *Nat. Neurosci.*, *14*, 1475–1479.

Sherman, M. T., Seth, A. K., Barrett, A. B., & Kanai, R. (2015). Prior expectations facilitate metacognition for perceptual decision. *Consciousness and Cognition*, *35*, 53–65.

Singh, K., & Scott, S. H. (2003). A motor learning strategy reflects neural circuitry for limb control. *Nature Neuroscience*, *6*, 399–403.

Smith, F. W., & Muckli, L. (2010). Nonstimulated early visual areas carry information about surrounding context. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 20099–20103.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: applications to dementia and amnesia. *Journal of Experimental Psychology. General*, *117*, 34–50.

Sorabji, R., & Aristote. (1972). *Aristotle on memory*. *Aristotle on memory*. London: Duckworth.

Spence, M. L., Dux, P. E., Arnold, D. H., Spence, M. L., Dux, P. E., & Arnold, D. H. (2015). Computations Underlying Confidence in Visual Perception. *Journal of Experimental Psychology : Human Perception and Performance*, *41*.

Spratling, M. W. (2008). Predictive coding as a model of biased competition in visual attention. *Vision Research*, *48*, 1391–408.

Spratling, M. W. (2008). Reconciling Predictive Coding and Biased Competition Models of Cortical Function. *Frontiers in Computational Neuroscience*, *2*, 8.

Spratling, M. W. (2016). A review of predictive coding algorithms. *Brain and Cognition*. doi:10.1016/j.bandc.2015.11.003

St. John-Saaltink, E., Utzerath, C., Kok, P., & Lau, H. C. (2015). Expectation Suppression in Early Visual Cortex Depends on Task Set, *10*, 1–14.

Stalnaker, T., Cooch, N. K., & Schoenbaum, G. (2015). What the orbitofrontal cortex does not do. *Nature Neuroscience*, *18*, 620–627.

Stankevicius, A., Huys, Q. J. M., Kalra, A., & Seriès, P. (2014). Optimism as a Prior Belief about the Probability of Future Reward. *PLoS Computational Biology*, *10*, 1–21.

Stefanics, G., Hangya, B., Hernádi, I., Winkler, I., Lakatos, P., & Ulbert, I. (2010). Phase entrainment of human delta oscillations can mediate the effects of expectation on reaction speed. *The Journal of Neuroscience*, *30*, 13578–85.

Stefanics, G., Kremlacek, J., & Czigler, I. (2014). Visual mismatch negativity: a predictive coding view. *Frontiers in Human Neuroscience*, *8*, 1–19.

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

Sterzer, P., Frith, C. D., & Petrovic, P. (2008). Believing is seeing : expectations alter visual awareness Neural basis for unique hues. *Current Biology*, *18*, R697–R698.

Summerfield, C., & de Lange, F. P. (2014). Expectation in perceptual decision making: neural and computational mechanisms. *Nature Reviews Neuroscience*. doi:10.1038/nrn3838

Summerfield, C., & Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends in Cognitive Sciences*, *13*, 403–9.

Summerfield, C., & Egner, T. (2016). Feature-Based Attention and Feature-Based Expectation. *Trends in Cognitive Sciences*, *xx*, 1–4.

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

Summerfield, C., & Mangels, J. a. (2006). Dissociable neural mechanisms for encoding predictable and unpredictable events. *Journal of Cognitive Neuroscience*, *18*, 1120–1132.

Summerfield, C., Monti, J. M. P., Trittschuch, E. H., Mesulam, M., & Egner, T. (2009). Neural repetition suppression reflects fulfilled perceptual expectations. *Nature Neuroscience*, *11*, 1.

Suzuki, M., & Gottlieb, J. (2013). Distinct neural mechanisms of distractor suppression in the frontal and parietal lobe. *Nature Neuroscience*, *16*, 98–104.

Sy, J. L., Guerin, S. a, Stegman, A., & Giesbrecht, B. (2014). Accurate expectancies diminish perceptual distraction during visual search. *Frontiers in Human Neuroscience*, *8*, 334.

Synofzik, M., Lindner, A., & Thier, P. (2008). The Cerebellum Updates Predictions about the Visual Consequences of One's Behavior. *Current Biology*, *18*, 814–818.

Theeuwes, J. (2010). Top-down and bottom-up control of visual selection. *Acta Psychologica*, *135*, 77–99.

Timmermans, B., Schilbach, L., Pasquali, a., & Cleeremans, a. (2012). Higher order thoughts in action: consciousness as an unconscious re-description process.

Philosophical Transactions of the Royal Society B: Biological Sciences, *367*, 1412–1423.

Ting, C.-C., Yu, C.-C., Maloney, L. T., & Wu., S.-W. (2015). Neural Mechanisms for Integrating Prior Knowledge and Likelihood in Value-Based Probabilistic Inference. *The Journal of Neuroscience*, *35*, 1792–1805.

Todorovic, A., van Ede, F., Maris, E., & de Lange, F. P. (2011). Prior expectation mediates neural adaptation to repeated sounds in the auditory cortex: an MEG study. *The Journal of Neuroscience*, *31*, 9118–23.

Trapp, S., & Bar, M. (2015). Prediction, context, and competition in visual recognition. *Annals of the New York Academy of Sciences*, *1339*, 190–198.

Umbach, V. J., Schwager, S., Frensch, P. a, & Gaschler, R. (2012). Does explicit expectation really affect preparation? *Frontiers in Psychology*, *3*, 378.

Van de Cruys, S., Evers, K., Van der Hallen, R., Van Eylen, L., Boets, B., De-Wit, L., & Wagemans, J. (2014). Precise minds in uncertain worlds: predictive coding in autism. *Psychological Review*, *121*, 649–75.

van Ede, F., Jensen, O., & Maris, E. (2010). Tactile expectation modulates pre-stimulus beta-band oscillations in human sensorimotor cortex. *NeuroImage*, *51*, 867–76.

van Gaal, S., & Lamme, V. A. F. (2012). Unconscious high-level information processing: implication for neurobiological theories of consciousness. *The Neuroscientist : A Review Journal Bringing Neurobiology, Neurology and Psychiatry*, *18*, 287–301.

van Kerkoerle, T., Self, M. W., Dagnino, B., Gariel-Mathis, M. -a., Poort, J., van der Togt, C., & Roelfsema, P. R. (2014). Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex. *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.1402773111

Vandenbroucke, A. R. E., Sligte, I. G., Barrett, A. B., Seth, A. K., Fahrenfort, J. J., & Lamme, V. A. F. (2014). Accurate metacognition for visual sensory memory representations. *Psychological Science*. doi:10.1177/0956797613516146

Vanrullen, R., Busch, N. a, Drewes, J., & Dubois, J. (2011). Ongoing EEG Phase as a Trial-by-Trial Predictor of Perceptual and Attentional Variability. *Frontiers in Psychology*, *2*, 60.

VanRullen, R., & Thorpe, S. J. (2001). The Time Course of Visual Processing: From Early Perception to Decision-Making. *Journal of Cognitive Neuroscience*, *13*, 454–461.

Verbruggen, F., & Logan, G. D. (2008). Response inhibition in the stop-signal paradigm. *Trends in Cognitive Sciences*, *12*, 418–24.

Vickers, D. (1970). Evidence for an accumulator model of psychophysical discrimination. *Ergonomics*, *13*, 37–58.

Vickers, D. (1979). *Decision processes in visual perception*. New York, NY: Academic Press.

Vickers, D., & Packer, J. (1982). Effects of alternating set for speed or accuracy on response time, accuracy and confidence in a unidimensional discrimination task. *Acta Psychologica*, *50*, 179–197.

Vickers, D., Smith, P., Burt, J., & Brown, M. (1985). Experimental paradigms emphasising state or process limitations: II effects on confidence. *Acta Psychologica*, *59*, 163–193.

Vilares, I., Howard, J. D., Fernandes, H. L., Gottfried, J. a., & Kording, K. P. (2012). Differential representations of prior and likelihood uncertainty in the human brain. *Current Biology*, *22*, 1641–1648.

Vlassova, A., Donkin, C., & Pearson, J. (2014). Unconscious information changes decision accuracy but not confidence. *Proceedings of the National Academy of Sciences*, *111*, 16214–16218.

von Stein, a, Chiang, C., & König, P. (2000). Top-down processing mediated by interareal synchronization. *Proceedings of the National Academy of Sciences of the United States of America*, *97*, 14748–53.

Vossel, S., Geng, J. J., & Fink, G. R. (2014). Dorsal and Ventral Attention Systems Distinct Neural Circuits but Collaborative Roles. *The Neuroscientist*, *20*, 150–159.

Vossel, S., Mathys, C., Daunizeau, J., Bauer, M., Driver, J., Friston, K. J., & Stephan, K. E. (2014). Spatial attention, precision, and Bayesian inference: a study of saccadic response speed. *Cerebral Cortex (New York, N.Y. : 1991)*, *24*, 1436–50.

Wacongne, C., Changeux, J. P., & Dehaene, S. (2012). A Neuronal Model of Predictive Coding Accounting for the Mismatch Negativity. *The Journal of Neuroscience*, *32*, 3665–3678.

Wacongne, C., Labyt, E., Van Wassenhove, V., Bekinschtein, T., Naccache, L., & Dehaene, S. (2011). Evidence for a hierarchy of predictions and prediction errors in human cortex. *Proceedings of the National Academy of Sciences*, *108*, 1–6.

Wallis, J. D. (2007). Orbitofrontal cortex and its contribution to decision-making. *Annual Review of Neuroscience*, *30*, 31–56.

Wei, X.-X., & Stocker, A. a. (2015). A Bayesian observer model constrained by efficient coding can explain “anti-Bayesian” percepts. *Nature Neuroscience*, *18*, 1509–17.

White, T. P., Engen, N. H., Sørensen, S., Overgaard, M., & Shergill, S. S. (2014). Uncertainty and confidence from the triple-network perspective: voxel-based meta-analyses. *Brain and Cognition*, *85*, 191–200.

Wierzchoń, M., Paulewicz, B., Asanowicz, D., Timmermans, B., & Cleeremans, A. (2014). Different subjective awareness measures demonstrate the influence of visual identification on perceptual awareness ratings. *Consciousness and Cognition*, *27C*, 109–120.

Wilimzig, C., & Fahle, M. (2008). Spatial attention increases performance but not subjective confidence in a discrimination task. *Journal of Vision*, *8*, 1–10.

Womelsdorf, T., & Fries, P. (2007). The role of neuronal synchronization in selective attention. *Current Opinion in Neurobiology*, *17*, 154–60.

Wong, K.-F. (2006). A Recurrent Network Mechanism of Time Integration in Perceptual Decisions. *The Journal of Neuroscience*, *26*, 1314–1328.

Wyart, V., Nobre, A. C., & Summerfield, C. (2012). Dissociable prior influences of signal probability and relevance on visual contrast sensitivity. *Proceedings of the National Academy of Sciences*, *109*, 6354–6354.

Wyart, V., & Tallon-Baudry, C. (2008). Neural dissociation between visual awareness and spatial attention. *The Journal of Neuroscience*, *28*, 2667–2679.

Wyart, V., & Tallon-Baudry, C. (2009). How ongoing fluctuations in human visual cortex predict perceptual awareness: baseline shift versus decision bias. *The Journal of Neuroscience*, *29*, 8715–25.

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making:

confidence and error monitoring. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*, 1310–21.

Yokoyama, O., Miura, N., Watanabe, J., Takemoto, A., Uchida, S., Sugiura, M., … Nakamura, K. (2010). Right frontopolar cortex activity correlates with reliability of retrospective rating of confidence in short-term recognition memory performance. *Neuroscience Research*, *68*, 199–206.

Yuille, A., & Kersten, D. (2006a). Vision as Bayesian Inference : Analysis by Synthesis ? Introduction : Perception as inference. *Trends in Cognitive Sciences*, *10*, 301–308.

Yuille, A., & Kersten, D. (2006b). Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, *10*, 301–308.

Zarnoth, P., & Sniezek, J. a. (1997). The Social Influence of Confidence in Group Decision Making, *366*, 345–366.

Zizlsperger, L., Sauvigny, T., & Haarmeier, T. (2012). Selective attention increases choice certainty in human decision making. *PLoS ONE*, *7*. doi:10.1371/journal.pone.0041136

Zumer, J. M., Scheeringa, R., Schoffelen, J.-M., Norris, D. G., & Jensen, O. (2014). Occipital Alpha Activity during Stimulus Processing Gates the Information Flow to Object-Selective Cortex. *PLoS Biology*, *12*, e1001965.

Zylberberg, A., Barttfeld, P., & Sigman, M. (2012). The construction of confidence in a perceptual decision. *Frontiers in Integrative Neuroscience*, *6*, 79.

Zylberberg, A., Roelfsema, P. R., & Sigman, M. (2014). Variance misperception explains illusions of confidence in simple perceptual decisions. *Consciousness and Cognition*, *27C*, 246–253.