University of Glasgow | Institute of Neuroscience & Psychology

Spatiotemporal neural correlates of confidence in perceptual decision making

Ana Sabina Gherman

BSc Psychology, MSc Brain Imaging

Submitted in fulfilment of the requirements for the Degree of Doctor of Philosophy

Institute of Neuroscience and Psychology

College of Medical, Veterinary and Life Sciences

University of Glasgow

September 2017

Abstract

In our interactions with the environment, we often make inferences based on noisy or incomplete perceptual information - for example, judging whether the person waving their hand in the distance is someone we know (as opposed to a stranger, greeting the person behind us). Such judgments are accompanied by a sense of confidence, that is, a degree of belief that we are correct, which ultimately determines how we act, adjust our subsequent decisions, or learn from errors. Neuroscience has only recently begun to characterise the representations of confidence in the animal and human brain, however the neural mechanisms and network dynamics supporting these representations are still unclear.

The current thesis presents empirical findings from three studies that sought to provide a more complete characterisation of confidence during perceptual decision making, using a combination of electrophysiological and neuroimaging methods. Specifically, **Study 1** (Chapter 2) investigated the temporal characteristics of confidence in relation to the perceptual decision. We recorded EEG measurements from human subjects during performance of a face vs. car categorisation task. On some trials, subjects were offered the possibility to opt out of the choice in exchange for a smaller but certain reward (relative to the reward obtained for correct choices), and the choice to use or decline this option reflected subjects' confidence in their perceptual judgment. Neural activity discriminating between high vs. low confidence trials could be observed peaking approximately 600 ms after stimulus onset. Importantly, the temporal profile of this activity resembled a ramp-like process of evidence accumulation towards a decision, with confidence being reflected in the rate of the accumulation. Our results are in line with the notion that neural representations of confidence may arise from the same process that supports decision formation.

Extending on these findings, in **Study 2** (Chapter 3) we asked whether rhythmic patterns within the EEG signals may offer additional insights into the neural representations of confidence. Using an exploratory analysis of data from Study 1, we identified confidence-discriminating oscillatory activity in the alpha and

beta frequency bands. This was most prominent over the sensorimotor electrodes contralateral to the motor effector that subjects used to indicate choice (i.e., right hand), consistent with a motor preparatory signal. Importantly however, the effect was transient in nature, peaking long before subjects could execute a response, and thus ruling out a direct link with overt motor behaviour. More intriguingly, the observed confidence effect appeared to overlap in time with the non-oscillatory representation of confidence identified in Study 1. In line with the view that motor systems track the evolution of the perceptual decision in preparation for impending action, results from Studies 1 and 2 open the possibility that confidence-related information may also be contained within these signals.

Finally, following on from our work in the first study, we next aimed to capitalise on the single-trial neural representations of confidence obtained with EEG, in order to identify potentially correlated activity with high spatial resolution. To this end, in **Study 3** (Chapter 4) we recorded simultaneous EEG and fMRI data while subjects performed a speeded motion discrimination task and rated their confidence on a trial-by-trial basis. Analysis of the EEG revealed a confidence-discriminating neural component which appeared prior to participants' overt choice and was spatiotemporally consistent with our results from the first study. Crucially, we showed that haemodynamic responses in the ventromedial prefrontal cortex (VMPFC) were uniquely explained by trial-to-trial fluctuations in these early confidence-related neural signals. Notably, this activation was additional to what could be explained by subjects' confidence ratings alone. We speculated that the VMPFC may support an early and/or automatic readout of perceptual confidence, potentially preceding explicit metacognitive appraisal.

Together, our results reveal novel insights into the neural representations of perceptual confidence in the human brain, and point to new research directions that may help further disentangle the neural dynamics supporting confidence and metacognition.

Table of Contents

Acknowledgments

Above all, I would like to thank my supervisor, Dr. Marios Philiastides, for his invaluable support and guidance throughout the most important years of my academic development. I am deeply grateful for having been part of your lab, and for the amazing learning opportunities working with you has opened. Thank you for your kindness, patience, and ever-uplifting spirit, and most of all, for never running out of the encouraging words I needed to complete this thesis. It made all the difference.

Thank you also to Frances Crabbe, for kindly sharing her MRI expertise, and offering much-needed assistance with data collection. I am also grateful to all the volunteers who resiliently endured the long hours of testing for the sake of science.

To all the wonderful people with whom I shared the colourful range of PhD-related experiences, from data analysis, experiments, conferences, to travels, food, and distractions. Jessy, Elsa, Andrea, Leon, Gabby, Essi, Ema, Filippo, Alex, Kevin, Kasia, Fei, and Steph - thank you for being the social side of my academic life.

To Henrique, thank you for being my most valued source of human interaction throughout the PhD journey.

Finally, I am immensely grateful to my parents and sister for their unconditional love and support, and to my mother in particular, who will never stop looking after me no matter how old I get.

List of Tables

List of Figures

Chapter 2

List of Publications

Gherman, S. and Philiastides, M.G., 2015. Neural representations of confidence emerge from the process of decision formation during perceptual choices. *Neuroimage*, *106*, pp.134-143.

Gherman, S. and Philiastides, M.G. (submitted). Human VMPFC encodes early signatures of confidence in perceptual decisions.

Author's Declaration

I declare that, except where explicit reference is made to the contribution of others, that this dissertation is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution.

Abbreviations

BOLD	Blood oxygen level dependent
dB	Decibel
EEG	Electroencephalography
fMRI	Functional magnetic resonance
GLM	General linear model
LIP	Lateral intraparietal
OFC	Orbitofrontal cortex
PFC	Prefrontal cortex
RLPFC	Rostrolateral prefrontal cortex
SEF	Supplementary eye field
SR	Sure reward
TMS	Transcranial magnetic stimulation
VMPFC	Ventromedial prefrontal cortex

Chapter 1. General Introduction

Every day we make judgments about perceptual aspects of our environment (i.e., perceptual decisions), on the basis of noisy or incomplete information. Such judgments are invariably accompanied by a sense of likelihood that we are correct, and we rely on these to optimally interact with the external world. Having access to an internal estimate of decision accuracy is essential in regulating adaptive behaviour in an uncertain world - our sense of confidence in a judgment can influence subsequent decisions and actions (Folke et al. 2016, Kepecs et al. 2008, Kiani and Shadlen 2009, Lak et al. 2014, van den Berg et al. 2016b), and support learning processes (Guggenmos et al. 2016, Lak et al. 2017, Daniel and Pollmann 2012). Over the past century, the topic of decision confidence has attracted considerable scientific interest, with recent years in particular seeing rapid progress in characterising its behavioural, computational, and neurobiological correlates, in both humans and animals. Nevertheless, the neuroscientific study of decision confidence is only in its infancy and many questions are yet to be addressed. In particular, the mechanisms by which confidence in a perceptual decision is formed in the human brain, and the network dynamics that support these processes, are unclear. The current chapter will summarise research that has focused on characterising the neural correlates of perceptual decision making and associated confidence, in humans and animals, and outline outstanding questions that motivated the current thesis.

Perceptual decision making: neural mechanisms

Animals

The term “perceptual decision” is used to refer to the process of committing to one of several potential alternatives (i.e., judgments or choices), based on an integration of sensory information (Heekeren et al. 2008). This process has been described in the framework of sequential sampling models, which postulate that a decision is formed via a noisy accumulation of sensory information over time,

with the decision terminating when an internal threshold has been reached (Usher and McClelland 2001, Ratcliff 1978, Smith and Ratcliff 2004). Strong support for such a mechanism comes primarily from non-human primate neurophysiological research (see Gold and Shadlen, 2007, for a review). In these studies, monkeys are trained to perform two-alternative forced choice tasks, such as the random-dot motion discrimination paradigm (Newsome and Pare, 1988) and express their choice by making a saccade towards a target. Single-cell recordings have revealed that upon stimulation, choice-selective neurons in frontal and parietal areas such as the frontal eye field (Kim and Shadlen 1999), superior colliculus (Horwitz and Newsome 1999), or lateral intraparietal area (Shadlen and Newsome 2001, Roitman and Shadlen 2002) exhibit a gradual increase in firing rates, which remains elevated and reaches a common level before a response is made. Importantly, the profile of this activity is modulated by the quality of sensory evidence, with stronger stimulus strength eliciting steeper accumulation rates. Additionally, it predicts monkeys' choice-related behaviour, with steeper buildup of activity resulting in faster and more accurate responses (Shadlen and Newsome 2001, Roitman and Shadlen 2002).

Humans

Perceptual decisions in the human brain appear to be supported by a similar mechanism of bounded evidence accumulation. Specifically, electrophysiological (Van Vugt et al. 2012, Philiastides and Sajda 2006, de Lange et al. 2013, Donner et al. 2009, Philiastides et al. 2014, Wyart et al. 2012, Polania et al. 2014) and neuroimaging (Liu and Pleskac 2011, Ploran et al. 2007, Heekeren et al. 2004, Krueger et al. 2017) work has revealed signals which resemble the dynamic patterns observed in single-unit recordings. One example is a recent EEG study (Philiastides et al. 2014) where subjects were asked to perform visual categorisations of face vs. car stimuli. Authors revealed ramp-like signals over centroparietal electrodes, the slope of which scaled positively with the strength of the stimulus and matched predictions from a sequential sampling model of decision making (i.e., the drift diffusion model; Ratcliff, 1978). The buildup rate of this activity was additionally predictive of subjects' choice accuracy on a

trial-by-trial basis. A similar centroparietal signal was observed by O'Connell et al. (2012), who showed that the buildup of activity predicted subjects' response time even when stimulus difficulty remained constant, consistent with decision-related activity that reflects internal noise in the decision process. Importantly, both studies showed that this activity was independent of motor preparation. Similar patterns have been observed across different tasks and sensory modalities (O'Connell et al. 2012, Kelly and O'Connell 2013, Murphy et al. 2015), pointing to a potentially domain-general decision signal.

Oscillatory neural signals also appear to reflect decision-related processes. Specifically, activity resembling a process of bounded evidence accumulation has been observed in the theta (Van Vugt et al. 2012) and gamma (Polania et al. 2014) frequency bands. Intriguingly, a few studies have found that decision-related activity can be observed in action-selective neural signals, as measured with MEG. Namely, when subjects express their perceptual choices via motor behaviour (e.g., button presses), a reduction of oscillatory activity in the alpha and beta bands (approximately ~8-30 Hz), can be observed over the contralateral motor cortex, following perceptual stimulation and prior to overt choice. Although typically associated with motor-related planning and preparation (Pfurtscheller and Lopes da Silva 1999), this activity nevertheless occurs long before a response is made, scales with accumulated evidence within upstream (sensory) regions (Donner et al. 2009), and its slope is modulated by stimulus strength (de Lange et al. 2013), consistent with a decision-related process. Interestingly, these signals can appear as early as the decision signals observed in the time domain (O'Connell et al. 2012). While there is strong empirical evidence that motor-preparatory activity is distinct from action-*independent* decision processes (Kelly and O'Connell 2013, Wyart et al. 2012, Filimon et al. 2013), this finding has supported the view that decision-related information may also be carried by motor systems in support of impending actions (Gold and Shadlen 2007, Gold and Shadlen 2000, Siegel et al. 2011).

Confidence in perceptual decision making

As the neural correlates of perceptual decisions are being uncovered, there has been growing interest in understanding how confidence in these decisions may arise and become available for metacognitive evaluation and report. The following sections provide a brief review of the empirical work aimed at characterising the neural basis of confidence in perceptual decisions.

Measuring confidence

The methods that have been used most commonly to obtain behavioural measures of confidence can broadly be categorised according to their explicit or implicit nature (see Kepecs and Mainen (2012) for a detailed review). Human experiments typically rely on explicit reports, whereby subjects provide confidence ratings upon making a task-related choice. These can be verbal reports, where subjects select from discrete categories (e.g., “High” vs. “Low”, Peters et al., 2017) or make use of scales (e.g., ranging from “Not at all confident” to “Totally confident”, Lebreton et al., 2015). Alternatively, and more commonly, subjects are asked to use numerical or visual analogue scales (Fleming et al. 2010, Festinger 1943, Baranski and Petrusic 1994, Hebart et al. 2016), where the lowest value typically indicates a guess.

Implicit measures of confidence require the experimental design to be constructed such that subject’s choices reflect confidence indirectly. One variant that has been used in research on rodents is the waiting-based method (Kepecs et al. 2008, Lak et al. 2014). Upon making a perceptual decision, subjects can choose to wait for a delayed reward (which is provided only for correct responses) or alternatively abort the trial to initiate a new one. In this paradigm, subjects’ willingness to wait for a reward is predictive of the likelihood of making a correct response, thus serving as a proxy for confidence. An alternative approach is the wagering technique, which requires subjects to choose between safer vs. riskier (but potentially more rewarding) options, the outcome of which depends on the accuracy of their (over or covert) decision (Middlebrooks and Sommer 2012, Kiani and Shadlen 2009). One variant of this

method is the “opt-out” task, used predominantly in the monkey literature (Kiani and Shadlen 2009, Odegaard et al. 2017, Komura et al. 2013). Subjects make perceptual discriminations which are rewarded for correct responses. Importantly, on some trials, in addition to the two stimulus alternatives, a third response option is available which allows subjects to opt out of the choice in exchange for a smaller but certain reward. The rationale behind this approach is that the choice to select or waive the sure reward option reflects the subjective belief that a judgment is correct. Indeed, studies employing this task show that subjects are more likely to be accurate on trials where the opt-out was offered and declined, compared to those in which it was not offered to begin with (Kiani and Shadlen 2009).

In humans, this method may provide an advantage over the classic rating task, in that subjects must use the internal evaluation of their judgment accuracy to maximise their rewards, thus serving as an incentive to accurately reveal this information (Persaud et al. 2007). A potential downside, however, is that opt-out behaviour can also be influenced by subjects’ aversion to risk (Fleming and Dolan 2010), which is not an issue in ratings tasks. An additional advantage of the rating tasks is the ability to obtain graded measures of confidence (as compared with binary values obtained with opt-out tasks), which may allow for more precise inferences about underlying neural representations.

Behavioural correlates and theoretical framework

Early studies investigating the behavioural properties of confidence have revealed close links with quantities known to influence, or reflect, the decision process. In particular, it is well-established that confidence tends to increase with the strength of sensory information (Peirce and Jastrow 1884, Festinger 1943, Baranski and Petrusic 1998). Additionally, confidence correlates with behavioural manifestations of the decision, such as choice accuracy and response time. Confident choices are more likely to be correct (Baranski and Petrusic 1998), and are associated with shorter response times (Baranski and Petrusic 1998, Festinger 1943, Vickers and Packer 1982). These observations reinforce the

idea that confidence is a fundamental aspect of the decision process, and have led to both implicit and explicit assumptions that confidence of a decision is based on the same process that underlies the decision (Vickers 1979, Kepecs et al. 2008, Hebart et al. 2016, Kiani and Shadlen 2009, Fetsch et al. 2014). There is however growing evidence that confidence can, in some instances, be dissociated from the decision process itself. Behaviourally, this is best reflected by incongruences between objective task performance and subjective evaluation of one's performance. For example, humans tend to be overconfident in their choices when stimulus strength is poor (and performance consequently lower), and conversely underestimate their performance when the task is easy (Baranski and Petrusic 1994, Baranski and Petrusic 1999, Zylberberg et al. 2014). Similarly, the ability to accurately estimate one's own performance (i.e., metacognitive ability) can vary across individuals (Fleming et al. 2010, Fleming et al. 2012), such that high performance on a task can be accompanied by near-chance performance on the metacognitive task. Theoretical frameworks accounting for such dissociations between decision and performance have suggested that confidence relies on, or can be influenced by, additional processes occurring after the decision (Moran et al. 2015, Yu et al. 2015, Pleskac and Busemeyer 2010, Baranski and Petrusic 1998). For example, the two-stage dynamic signal detection (2DSD) (Pleskac and Busemeyer 2010), a type of sequential sampling model, posits that the process of evidence accumulation leading to a decision continues to develop after the choice to inform confidence. Such a view is additionally supported by the observation that decisions can be promptly followed by changes of mind (Resulaj et al. 2009, van den Berg et al. 2016a), suggestive of additional processing beyond the initial choice.

Neural correlates

Animals

As pointed out in the previous sections, the ability to access information about one's performance is not limited to humans, and can also be observed in other species. Indeed, rodents and non-human primates appear to use internal

estimates of accuracy to maximise rewards (Kepecs and Mainen 2012, Middlebrooks and Sommer 2012, Kiani and Shadlen 2009, Lak et al. 2014). This discovery has been critical for characterising confidence-related processes at the neural level. Single-unit recordings in the animal brain make it possible to observe confidence-related neural activity with both high temporal and high spatial precision, whereas pharmacological inactivation studies can additionally reveal causal links with behaviour.

An important insight into the possible neural mechanisms underlying confidence comes from a seminal study by Kiani and Shadlen (2009). In their experiment, rhesus monkeys were trained to perform a random-dot motion discrimination task, whereby confidence was measured by means of an opt-out method (see previous sections). Choice-selective neurons within the lateral intraparietal (LIP) cortex exhibited choice-related buildup in firing rates, consistent with the process of evidence accumulation observed previously in this region. More importantly however, this activity also predicted confidence in the decision, i.e., whether the monkey would select or decline the sure reward option. Specifically, confident trials were characterised by a higher buildup rate, with activity reaching higher magnitudes prior to choice. Overall, these findings indicate that confidence-related information may emerge from the decision process itself, i.e., is encoded in the neural activity that supports it. A similar observation was made by Middlebrooks and Sommer (2012). They identified neurons in the supplementary eye field exhibiting differential activity for both choice (correct vs. error) and confidence (high vs. low), with this activity showing considerable temporal overlap. As will be discussed in the following section, these observations raise the possibility that a similar mechanism might underlie decisions in the human brain.

Two recent studies have pointed out that representations of confidence may occur independently of the decision process. For example, pharmacological inactivation of the OFC was shown to affect rats' ability to optimally wait for a performance-dependent reward, indicating disrupted internal estimates of decision accuracy and/or outcome. Despite this effect on confidence, task performance per se remained unhindered (Lak et al. 2014). Similarly, Komura et

al. (2013) showed that pharmacological inactivation of the monkey pulvinar (a region of the visual thalamus) increased the number of times monkeys made an opt-out choice (suggesting lower confidence), without affecting performance on the perceptual task. These studies point to a possible dissociation between regions that carry neural representations of confidence vs. choice.

Interestingly, representations of confidence have also been identified in regions of the brain involved in reward and learning. Neurons in the orbitofrontal cortex (OFC), a region implicated in decision making and reward processing (Wallis 2007), have been shown to carry confidence-related information during an olfactory categorisation task. Similarly, midbrain dopamine neurons, which are known to play a role in reward prediction and learning, also appear to encode a form of confidence. De Lafuente and Romo (2011) found that dopamine firing rates in the monkey brain were modulated by stimulus strength during correct detections of a vibrotactile stimulus, but not during missed trials, suggesting activity here was linked to the monkey's subjective experience (as opposed to objective stimulus properties). Extending these findings, Lak et al. (2017) showed that learning signals within dopamine neurons appeared to incorporate a measure of objective confidence (as estimated by an extended reinforcement learning model). Interestingly, these signals were observed prior to overt choices, leading authors to speculate that these reflect the evolving decision and could potentially influence impending choices.

Overall, findings from animal research suggest that the brain may carry multiple representations of confidence, potentially supporting different cognitive processes and behaviours. In regions such as the LIP and SEF, a form of confidence may emerge from the decision process, whereas regions such as the pulvinar and OFC appear to encode confidence separately from the decision. Bayesian theories of neural computation (Knill and Pouget 2004) suggest that the brain represents perceptual decisions in the form of probability distributions. Within this framework, confidence information is naturally present in the decision-related neural code (Meyniel et al. 2015, Pouget et al. 2016), in line with the role of LIP or SEF in encoding both choice and confidence. In a similar

line of reasoning, one mechanistic account of confidence proposes a framework by which confidence-related information emerging from the decision process is read-out by higher-order monitoring networks (Insabato et al. 2010), and it has been suggested that frontal regions, such as the OFC in the rat brain, may be likely candidates for such a role (Pouget et al. 2016, Lak et al. 2014).

Humans

Temporal correlates. In humans, the neural substrates of decision confidence have been explored using primarily non-invasive methods such as electroencephalography (EEG), magnetoencephalography (MEG), functional magnetic resonance imaging (fMRI), and transcranial magnetic stimulation (TMS). The millisecond temporal resolution of EEG and MEG provides a valuable tool for temporally characterising confidence-related processes, which in turn can help uncover underlying neural mechanisms. Nevertheless, only a limited number of studies have investigated the temporal correlates of confidence in human subjects. Of these, some have focused on events occurring after subjects have committed to a response, showing that signals that follow termination of the overt choice (i.e., motor response) reflect metacognitive processes (Murphy et al. 2015, Boldt and Yeung 2015). For example, Boldt and Yeung (2015) investigated the relationship between post-decision error-detection and confidence processing, bringing evidence for a common neural signature for the two (i.e., the classic error-positivity, or Pe, evoked component). Interestingly however, they also show that the amplitude of the stimulus-locked evoked component P300, which has been linked to evidence accumulation towards a decision (Twomey et al. 2015, Murphy et al. 2015), was modulated by reported confidence. While an interesting observation, the question of how this signal may relate to the decision process itself was not explicitly addressed here. Two studies have explicitly investigated the temporal characteristics of decision confidence relative to the decision. Zizlsperger et al. (2014) recorded scalp EEG from subjects during performance of a random-dot motion categorization task. They showed that ERP signals discriminated between levels of self-reported confidence as early as 300 ms following stimulus onset. This effect, which was

observed over occipitoparietal electrodes, was closely preceded by a neural representation of stimulus difficulty with similar topography, leading authors to suggest that the perceptual decision and confidence-related processes may overlap in time and share a neural substrate. Finally, a recent study (Peters et al. 2017) recorded intracranial EEG during a face vs. house categorization task. Subjects' choices revealed that sensory evidence was used differently for making a choice vs. reporting confidence, indicating a dissociation between the two processes. Interestingly, a dissociation between confidence and the decision could also be observed at the neural level, as reflected by stronger and earlier choice-related discrimination of neural signals. However, the spatial profile of this early choice-related activity (i.e., seen primarily over occipital regions) makes it unclear whether this may have reflected the decision process itself, or rather, an earlier process related to sensory evidence encoding, a distinction supported by monkey neurophysiology and human fMRI experiments (Heekeren et al. 2004, Gold and Shadlen 2007).

Spatial correlates. Similarly to animal work, studies in human subjects have revealed distributed networks that appear to hold neural representations of confidence, with regions of the prefrontal cortex (PFC) being most frequently observed in fMRI experiments (Hilgenstock et al. 2014, Rolls et al. 2010b, Fleming et al. 2012, Lau and Passingham 2006, Fleck et al. 2006, Heereman et al. 2015). The anterior portion of the PFC, in particular, appears to play a role in metacognitive evaluation of perceptual decisions (Baird et al. 2013, Fleming et al. 2010, Fleming et al. 2012). One fMRI study explicitly demonstrating the role of the anterior PFC in metacognition was conducted by Fleming et al. (2012). Participants performed face vs. house categorisations and were asked to rate their confidence after each choice. Blood oxygen level-dependent (BOLD) activity in the rostrolateral prefrontal cortex (RLPFC) correlated with confidence at the time of rating, and was enhanced during confidence rating compared to a control task. Importantly, the strength of the relationship between RLPFC activation and confidence reports was predictive of subjects' metacognitive ability, thus implicating this region in metacognitive processes. In support of this finding, it has also been shown that metacognitive ability correlates with macro- (Fleming et al. 2010) and microstructure (Allen et al. 2017) of the anterior PFC,

whereas damage to this region appears to impair metacognitive ability in perceptual decision making (Fleming et al. 2014)(though a recent study also showed improvement in metacognitive ability with temporary TMS disruption of activity in this region). Interestingly, correlates of perceptual confidence have also been detected in the striatum, a structure involved in reward processing. Specifically, Hebart et al. (2016) reported a positive correlation with reported confidence in the ventral portion of this region during a random-dot motion discrimination task. They speculate confidence-related striatal activation could represent implicit reward signals, which may serve to drive learning.

Overall, humans studies have focused predominantly on characterising confidence as a metacognitive process. However, as shown in the previous sections, confidence-related information can be observed earlier, near the time of the decision itself, and prior to overt commitment to choice or explicit metacognitive evaluation (Kiani and Shadlen 2009, Zizlsperger et al. 2014, Middlebrooks and Sommer 2012). Moreover, there is growing support for the idea that confidence processing is supported by hierarchical architectures relying on integration of confidence-related information by higher-order networks (Insabato et al. 2010, De Martino et al. 2013), and involving post-decisional processes (Maniscalco and Lau 2016, Pleskac and Busemeyer 2010, Fleming and Daw 2017, Yu et al. 2015, Moran et al. 2015, Resulaj et al. 2009), thus allowing the introduction of additional noise or changes in confidence-related signals prior to metacognitive report. In support of this view, one fMRI experiment that has investigated the neural correlates of confidence during value-based choices (De Martino et al. 2013) found that confidence emerging from a value-based decision process was encoded the same region that supported the decision (i.e., the ventromedial prefrontal cortex (VMPFC). Importantly, they showed that the rostrolateral PFC appeared to encode a noisy readout of this quantity in support of metacognitive report.

Overall, it becomes clear that, to understand the neural underpinnings of these complex network dynamics involved in confidence processing, it is necessary to begin characterising confidence-related quantities with both high-temporal and high-spatial precision.

Simultaneous EEG/fMRI. To date, no known studies have simultaneously investigated the spatiotemporal correlates of decision confidence in humans. Using advanced methods for the analysis of EEG signals, it is possible extract time-resolved single-trial measures representing cognitive events of interest, which can then be spatially characterised with fMRI. In particular, single-trial multivariate analysis of the EEG (Sajda et al. 2009) differs from conventional ERP-averaging approaches in that it preserves trial-to-trial variability of the neural response, which may hold valuable information about underlying neural activity. This method relies on simultaneously integrating information across a large number of sensors, and on using this information to identify EEG components that optimally discriminate between the conditions of interest. As such, signal quality can be improved whilst simultaneously preserving temporal information that would otherwise be lost through averaging across trials. EEG data alone cannot however provide precise spatial information about neural activity. To overcome this limitation, recent advances in neuroimaging methods have been developed which make possible the simultaneous acquisition of EEG and fMRI measurements, and these are becoming more widely used in the study of decision making (Pisauro et al. 2017, Goldman et al. 2009, Fouragnan et al. 2015). Combined with the single-trial EEG analysis techniques, it is possible to characterise neural signals of interest with higher precision and spatiotemporal accuracy than allowed by either method alone. Namely, the single-trial variability in EEG components of interest can be used to detect functionally correlated activity in the fMRI BOLD signal. Applied to the study of confidence, this method makes it possible to capitalise on endogenous (i.e., neural) signals associated with confidence, and expose potential latent states that might not be captured by behavioural reports alone.

Aims of the thesis

As this chapter has highlighted, there is overall a growing body of research uncovering the neural correlates of decision confidence. Nevertheless, several questions merit additional consideration, some of which are addressed in the current thesis. Firstly, as presented earlier, empirical work in non-human primates suggests that confidence-related information may become available

early on in the decision process, and potentially encoded in the decision process itself. The possibility that such a mechanism might underlie perceptual confidence in the human brain has not yet been explicitly assessed. This question motivated our first study, which will be presented in Chapter 2. In short, we collected EEG measurements from human subjects during performance of a face vs. car visual categorisation task. Using a single-trial multivariate analysis of the EEG, we found that neural signals discriminating between high and low confidence displayed a temporal pattern consistent with a process of decision-related evidence accumulation. We showed that confidence was reflected in the rate of this buildup, in line with the notion that confidence-related information may be represented in the same neural process that supports the decision.

Our second study, which extended this work, is presented in Chapter 3. As highlighted above, rhythmic neural activity has been shown to contain information about the ongoing decision process, offering insights into the underlying neural mechanisms of decision making which *surpass* the information obtained from time-domain analyses. We thus asked whether such signals may also hold information about the confidence in the perceptual decision. Using data from our first study, we adopted an exploratory approach whereby we sought to characterise neural representations of confidence in the frequency domain.

Finally, Chapter 4 presents the third and final study, in which we aimed to capitalise on the trial-by-trial variability in the time-resolved, endogenous markers of confidence identified with EEG, to identify potentially correlated activation in the fMRI data. To this end we collected simultaneous EEG and fMRI recordings while subjects performed a random-dot motion discrimination task and rated their confidence on a trial-by-trial basis. The primary goal of this approach was to characterise confidence-related signals with higher spatiotemporal precision than permitted by either method in isolation, and importantly, to obtain a more accurate representation of early confidence signals (i.e., occurring near the time of the decision and prior to explicit metacognitive evaluation) than has so far been possible in human studies.

Chapter 2. Neural representations of confidence emerge from the process of decision formation during perceptual choices

Summary

Choice confidence represents the degree of belief one's actions are likely to be correct or rewarding and plays a critical role in optimising our decisions. Despite progress in understanding the neurobiology of human perceptual decision-making, little is known about the representation of confidence. Importantly, it remains unclear whether confidence forms an integral part of the decision process itself or represents a purely post-decisional signal. To address this issue we employed a paradigm whereby on some trials, prior to indicating their decision, participants could opt-out of the task for a small but certain reward. This manipulation captured participants' confidence on individual trials and allowed us to discriminate between electroencephalographic signals associated with certain-vs-uncertain trials. Discrimination increased gradually and peaked well before participants indicated their choice. These signals exhibited a temporal profile consistent with a process of evidence accumulation, culminating at time of peak discrimination. Moreover, trial-by-trial fluctuations in the accumulation rate of nominally identical stimuli were predictive of participants' likelihood to opt-out of the task, suggesting confidence emerges from the decision process itself and is computed continuously as the process unfolds. Correspondingly, source reconstruction placed these signals in regions previously implicated in decision making, within the prefrontal and parietal cortices. Crucially, control analyses ensured that these results could not be explained by stimulus difficulty or changes in attention.

Introduction

Imagine running in the park on a rainy day, trying to discern whether the person across the lawn is an old friend. The decision to keep concentrating on your

stride or change directions to go greet them depends on your level of confidence that it is really them. Choice confidence is crucial not only for such mundane tasks, but also for more biologically and socially complex situations. It provides a probabilistic assessment of expected outcome and can play a key role in how we adjust in ever-changing environments, learn from trial and error, make better predictions, and plan future actions.

In recent years, systems and cognitive neuroscience have begun to examine the neural correlates underlying perceptual decision making. As a result, many monkey neurophysiology (Gold and Shadlen 2007, Kim and Shadlen 1999, Mazurek et al. 2003, Newsome et al. 1989, Shadlen et al. 1996, Shadlen and Newsome 2001), human neuroimaging (Heekeren et al. 2004, Heekeren et al. 2006, Heekeren et al. 2008, Ho et al. 2009, Ploran et al. 2007, Tosoni et al. 2008, Cheadle et al. 2014), and human electrophysiology (de Lange et al. 2010, Donner et al. 2009, Donner et al. 2007, Philiastides et al. 2006, Philiastides and Sajda 2006, Ratcliff et al. 2009, O'Connell et al. 2012, Wyart et al. 2012) experiments have provided converging support that perceptual decisions are characterised by a noisy temporal accumulation of sensory evidence which culminates when an observer commits to a choice. Despite this progress, however, it remains unclear how confidence is represented in the human brain and what its relationship is with the decision process itself.

Current theoretical and experimental accounts have regarded confidence as a metacognitive event that relies on new information arriving beyond the decision point (Fleming et al. 2012, Pleskac and Busemeyer 2010, Yeung and Summerfield 2012). Conversely, little has been done in the way of exploring whether confidence might emerge earlier in the decision process and before one commits to a choice. Evidence for the latter has recently emerged from a limited number of animal studies (Shadlen and Kiani 2013, Kiani and Shadlen 2009, Middlebrooks and Sommer 2012), proposing that choice confidence in perceptual judgments might be an inherent property of the decision process itself and that the same neural generators involved in evidence accumulation also encode choice confidence. To date, it remains unclear whether confidence forms an integral part of the decision process itself and whether it emerges from the same neural

generators involved in accumulating evidence for the decision. Similarly, it is unknown whether confidence is reflected in the rate of evidence accumulation itself.

To address these open questions, we collected electroencephalography (EEG) data during a binary, delayed-response, task in which correct responses were rewarded with monetary incentives. Importantly, on a random half of trials and after forming a decision, participants were given the option to opt out of the task for a smaller but sure reward (a form of post-decision wager; Kiani and Shadlen, 2009). We expected participants to waive the sure reward when they were certain of their choice, and select it otherwise. This in turn allowed us to use a multivariate single-trial classifier to discriminate between certain-vs-uncertain trials to identify the temporal characteristics of the neural correlates of choice confidence. Importantly, additional control analyses were carried out to ensure that confidence-related effects could not be explained by stimulus difficulty or trial-by-trial changes in attention.

Materials and Methods

Participants. Nineteen subjects (7 males) aged between 18-36 years (mean = 23.4 years) participated in the experiment. All had normal or corrected-to-normal vision and reported no history of neurological problems. Written informed consent was obtained in accordance with the School of Psychology Ethics Committee at the University of Nottingham.

Stimuli and task. Stimuli consisted of 20 face (face database, Max Planck Institute for Biological Cybernetics, Tuebingen, Germany) (Troje and Bulthoff 1996) and 20 car greyscale images obtained from the web (size 500×500 pixels, 8-bits/pixel). Spatial frequency, contrast, and luminance were equalised across all images, and the magnitude spectrum of each image was adjusted to the average magnitude spectrum of all images. We manipulated the phase spectrum of the images to obtain noisy stimuli of varying levels of sensory evidence (i.e. we manipulated the percentage phase coherence of our images) (Dakin et al.

2002). Stimuli were presented centrally on a plain grey background on a computer screen using PsychoPy software (Peirce 2007). The display was situated 1m away from the subject, with each stimulus subtending approximately 8 × 8 degrees of visual angle.

We used a training session prior to the main task to identify subject-specific phase coherence values for the stimuli used in the main task. Specifically, during training subjects were required to perform a simple speeded face vs. car categorisations over a total of 600 trials, using images with 7 different phase coherence values (27.5-42.5%, in increments of 2.5%). Each image was presented for 0.1 s and subjects were allowed a maximum of 1.25 s to make a response. The response was followed by an inter-trial interval, randomised between .75-1.5 s. There were an equal number of face and car stimuli, and these were presented in random order. Based on performance during this session, we selected three subject-specific phase coherence levels for the main task (henceforth referred to as Low, Medium, and High), which spanned psychophysical threshold (in the range 60-80% accuracy).

For the main experiment, subjects performed face vs. car categorisations during a delayed-response, post-decision wagering paradigm designed to discriminate between certain and uncertain trials (Fig. 2.1A). Importantly, on a random half of the trials, subjects were offered the option to opt-out of the task for a smaller (relative to a correct response) but sure reward (SR). This manipulation encouraged subjects to select the SR option on low confidence trials (Kiani and Shadlen 2009). Responses were rewarded with points (correct = 10 points, incorrect = 0 points, SR choice = 8 points). The total number of points collected was translated into a monetary payment at the end of the experiment. Each trial began with a face or car stimulus presented for 0.1s at one of the three possible sensory evidence levels. Stimulus presentation was followed by a forced delay (i.e., the decision time) randomised between 0.9-1.4s. This delay was introduced prior to revealing whether participants could opt-out of the task, to ensure they formed a decision on every trial. Next, a visual response cue (1s) informed participants whether or not the SR option would be available – this was

indicated by a green or red fixation cross, respectively. In addition, the letters “F” (for face) and “C” (for car) where positioned randomly to the left and right of the central fixation cross to indicate the mapping between stimulus and motor effectors (right index and ring fingers, respectively). The latter manipulation aimed at separating the decision process from motor planning and execution. Subjects indicated their choice by pressing one of three buttons on a response box (LEFT/RIGHT for a stimulus choice, MIDDLE for the SR). They were instructed to respond after the response cue was removed from the screen. A response was followed by an inter-trial interval randomised in the range 1-1.5 s. Overall subjects performed 480 trials, divided into two blocks of 240 trials each.

EEG data acquisition. We recorded EEG data during performance of the main task, in an electrostatically shielded room, using a DBPA-1 digital amplifier (Sensorium Inc., VT, USA), at a sampling rate of 1000Hz. We used 117 Ag/AgCl scalp electrodes and three periocular electrodes placed below the left eye and at the left and right outer canthi. Additionally, a chin electrode was used as ground. All channels were referenced to the left mastoid. Input impedance was adjusted to <50kOhm. To obtain accurate event onset times we placed a photodiode on the monitor to detect the onset of the stimuli. An external response device was used to collect response times. Both signals were collected on two external channels on the EEG amplifiers to ensure synchronization with the EEG data.

EEG data pre-processing. We applied a 0.5-100Hz band-pass filter to the data to remove slow DC drifts and high frequency noise. These filters were applied non-causally (using MATLAB “filtfilt”) to avoid phase related distortions. Additionally, we re-referenced our data to the average of all electrodes. To remove eye movement artefacts, participants performed an eye movement calibration task prior to the main experiment, during which they were instructed to blink repeatedly several times while a central fixation cross was displayed in the centre of the computer screen, and to make lateral and vertical saccades according to the position of the fixation cross. We recorded the timing of these visual cues and used principal component analysis to identify linear components

associated with blinks and saccades, which were then removed from the EEG data (Parra et al. 2005). Finally, we baseline corrected our EEG data, with the baseline interval defined as the 100ms prior to stimulus onset.

Single trial EEG analysis. To identify confidence-related activity in the neural data, we used a single-trial multivariate discriminant analysis (Parra et al. 2002, Parra et al. 2005) to estimate linear spatial weightings of the EEG sensors, which discriminated between certain (SR Waived) and uncertain (SR Selected) trials. We applied our technique to discriminate between the two groups of trials at various time points, in the time range between 100 ms prior to, and 1000 ms following the presentation of the visual stimulus (i.e. during the decision phase of the trial). For each participant we estimated, within short pre-defined time windows of interest, a projection in the multidimensional EEG space (i.e. a spatial filter) that maximally discriminated between the two conditions on stimulus-locked data (Eq. 1). Unlike conventional, univariate, trial-average event-related potential analysis, our multivariate approach is designed to spatially integrate information across the multidimensional sensor space, rather than across trials, to increase signal-to-noise ratio while preserving single-trial information.

Specifically, our method aimed to identify a one-dimensional 'discriminating component', $\boldsymbol{y}(t)$, by integrating information across all *D* electrodes, which maximally discriminated between the two trial groups. We use the term 'component' instead of 'source' to make it clear that this is a projection of all the activity correlated with the underlying source. We did this by applying a weighting vector $\boldsymbol{w}$ (i.e. a spatial filter) to our multidimensional EEG data ($\boldsymbol{x}(t)$), as summarised in the equation below:

$$\boldsymbol{y}(t) = \boldsymbol{w}^T \boldsymbol{x}(t) = \sum_{i=1}^{D} w_i x_i(t) \quad (1)$$

We used logistic regression and a reweighted least squares algorithm to learn the optimal discriminating spatial weighting vector $\boldsymbol{w}$ (Jordan and Jacobs 1994). We used this approach to identify a $\boldsymbol{w}$ for several short pre-defined training windows

centred at various latencies across our epoch of interest. Specifically, we used a 60 ms training window and stimulus-locked onset times varying from 100 ms before until 1000 ms after the stimulus, in increments of 10ms. The spatial filters ($\boldsymbol{w}$) obtained this way applied to an individual trial produce a measurement of the component amplitude for that trial. In separating the two groups of trials the discriminator was designed to map the component amplitudes for one condition to positive values and those of the other condition to negative values; note that this mapping was arbitrary. Here, we mapped the high confidence (SR Waived) trials to positive values and the low confidence (SR Selected) trials to negative values.

We quantified the performance of the discriminator for each time window using the area under a receiver operating characteristic (ROC) curve, referred to as an *Az*-value, using a leave-one-out procedure (Duda et al. 2001). To assess the significance of the discriminator we used a bootstrapping technique where we performed the leave-one-out test after randomising the trial labels. We repeated this randomization procedure 1000 times to produce a probability distribution for *Az*, and estimated the *Az* leading to a significance level of $p<0.01$.

To visualize the profile of the discriminating component, $\boldsymbol{y}$, across individual trials, we also constructed discriminant component maps (see Fig. 2.2C for an example). To do so we applied the spatial weighting vector $\boldsymbol{w}$ of the window that resulted in the highest discrimination performance between SR Waived vs. SR Selected trials, across an extended time range (100 ms before until 1000 ms after the stimulus). Each row of one such discriminant component map represents a single trial across time. We also sorted trials (i.e., the rows of these maps) based on the amplitude of the discriminating component in the time window of maximum discrimination. We also used this approach to compute the temporal profile of the discriminating component, $\boldsymbol{y}$, along the sensory evidence dimension to look for evidence of a gradual build-up of activity leading up to the point of maximum discrimination and to extract single-trial slopes of this accumulating activity. Slopes were computed using linear regression between

the onset- and peak times of the accumulating activity extracted from individual participants. Specifically, we extracted subject-specific accumulation onset-times by selecting (through visual inspection) the time point at which the discriminating component activity began to rise in a systematic fashion after an initial dip in the data following any early (non-discriminative) evoked responses present in the data (as seen in Fig. 2.4A). Peak accumulation times were selected as the time points of maximum discrimination across individual participants. To justify our choice for a linear model, we fit three additional models (exponential, logarithmic and power-law) to the individual subject accumulation patterns, using the same onset and peak accumulation times. We compared the goodness of fit to the data (mean square error) and found that the linear model provided the best fit to the accumulating activity, across all levels of sensory evidence.

Given the linearity of our model we also computed scalp projections of the discriminating components resulting from Eq. 1 by estimating a forward model for each component:

$$\mathbf{a} = \frac{\boldsymbol{X}\,\boldsymbol{y}}{\boldsymbol{y}^T\boldsymbol{y}} \tag{2}$$

where the EEG data ($\boldsymbol{X}$) and discriminating components ($\boldsymbol{y}$) are now in a matrix and vector notation, respectively, for convenience (i.e., both $\boldsymbol{X}$ and $\boldsymbol{y}$ now contain a time dimension). Equation 2 describes the electrical coupling of the discriminating component $\boldsymbol{y}$ that explains most of the activity in $\boldsymbol{X}$ (refer to Parra et al. 2002 for a detailed derivation of **a**). Strong coupling indicates low attenuation of the component $\boldsymbol{y}$ and can be visualised as the intensity of vector **a**. We used these scalp projections as a means of localizing the underlying neuronal sources (see next section).

Distributed source reconstruction. To spatially localize the resultant discriminating component activity related to choice confidence we used a distributed source reconstruction approach based on empirical Bayes (Friston et al. 2008) as implemented in SPM8 (http://www.fil.ion.ucl.ac.uk/spm/). The

method allows for an automatic selection of multiple cortical sources with compact spatial support that are specified in terms of empirical priors, while the inversion scheme allows for a sparse solution for distributed sources (refer to Friston et al., 2008, for details). We used a three-sphere head model, which comprised of three concentric meshes corresponding to the scalp, the skull and the cortex. The electrode locations were co-registered to the meshes using fiducials in both spaces and the head shape of the average MNI brain.

To compute the electrode activity to be projected onto these locations, we applied Eq. 2 to extract, at each time point, the scalp activity that was correlated with the confidence discriminating component $\boldsymbol{y}$ estimated during peak discriminator performance (i.e. we computed a forward model indexed by time, **a**(t)). We estimated **a**(t) in 1 ms data increments in the time range between 300 and 880 ms after stimulus onset (i.e. around the peak discrimination time).

Analysis of neural data. We used different logistic regressions to examine how neural activity correlated with participants' behavioural performance. To factor out the effect of task difficulty in our analyses, we first z-scored, at each level of sensory evidence separately, both the single-trial confidence component amplitudes (i.e., $\boldsymbol{y}$ at the end of the accumulation process) and the single-trial slopes of the accumulating activity itself (Acc. Slopes). Subsequently, we proceeded to perform different regression analyses on these trial-to-trial residual fluctuations (i.e., deviations from mean $\boldsymbol{y}$ and Acc. Slopes). Regression analyses were performed separately for each subject.

To assess how the fluctuations in discriminant component amplitude $\boldsymbol{y}$ (estimated from discriminating certain vs uncertain trials) influenced participants' likelihood of waiving the Sure Reward (SR), on trials where this option was available, we performed the following regression analysis:

$$P_{SR\,Waived} = [1 + e^{-(\beta_0 + \beta_1 y)}]^{-1} \quad (3)$$

We expected a positive correlation between the two quantities (as larger $\boldsymbol{y}$ amplitudes are expected to reflect more confident trials), and thus we tested whether the regression coefficients resulting across subjects ($\boldsymbol{\beta_1}$s in Eq. 3*)* came from a distribution with mean larger than zero (using a one-tailed t-test). We also repeated this analysis for each level of sensory evidence separately and tested whether $\boldsymbol{y}$ remained a significant predictor of participants' likelihood to waive the SR in each of the three levels. Moreover, we tested for differences in explanatory power across the three levels by comparing the resulting regression coefficients (using one-tailed paired t-test).

To assess how the slope of the accumulating activity influenced behavioural performance, we used the same rationale as with the previous analysis. Specifically, we used the accumulation slopes as a predictor for the probability of waiving the SR, on trials where this option was available:

$$P_{SR\,Waived} = [1 + \ e^{-(\beta_0 + \beta_1 \text{ Acc. Slopes})}]^{-1} \quad (4)$$

We hypothesised that, if confidence is an inherent property of the accumulation process itself, then accumulation slopes would be positively correlated with the probability of waiving the SR (i.e., $\boldsymbol{\beta_1}$>0), and we performed a one-tailed t-test to formally test for this hypothesis.

Next, we investigated whether accumulation slopes provided additional explanatory power for the probability of waiving the SR than what was already conferred by the discriminant component amplitude $\boldsymbol{y}$ (i.e. whether a significant positive correlation with accumulation slopes would still be present if the discriminant component amplitude $\boldsymbol{y}$ was included as an additional predictor in the regression):

$$P_{SR\,Waived} = [1 + \ e^{-(\beta_0 + \beta_1\, y + \beta_2 \text{ Acc. Slopes})}]^{-1} \quad (5)$$

As before, we a performed a one-tailed t-test to assess whether regression coefficients for accumulation slopes ($\boldsymbol{\beta_2}$s in Eq. 5) came from a distribution with mean larger than zero.

To rule out the possibility that confidence effects are driven by changes in attention across trials we included two additional predictors in the previous regression model, corresponding to two well-known neural signatures of attention; 1) pre-stimulus EEG power in the α band ($\boldsymbol{\alpha_{prestim}}$), which was linked to top-down control of attention (Wyart and Tallon-Baudry 2009) and was shown to correlate with visual discrimination performance (Thut et al. 2006, van Dijk et al. 2008), resulting from the analysis described in the next section and 2) an evoked component appearing 220 ms post-stimulus ($\boldsymbol{y_{220}}$), which was shown (in the same task used here) to index allocation of attentional resources required for the decision (Philiastides et al. 2006), and was localized in areas of the frontoparietal attention network (Philiastides and Sajda 2007).

$$P_{SR\ Waived} = [1 + e^{-(\beta_0 + \beta_1 y + \beta_2 \text{ Acc. Slopes} + \beta_3 \alpha_{prestim} + \beta_4 y_{220})}]^{-1} \quad (6)$$

We expected the fluctuations associated with confidence in both discriminant component amplitude $\boldsymbol{y}$ and accumulation slopes to remain a significant positive predictor of the likelihood of waiving the SR and thus we tested whether the resulting regression coefficients across subjects ($\boldsymbol{\beta_1}$s and $\boldsymbol{\beta_2}s$ in Eq. 6*)* came from a distribution with mean larger than zero (using a one-tailed t-test).

Single-trial power analysis. Pre-stimulus alpha power was obtained using a wavelet transform as in (Tallon-Baudry et al. 1996, Mazaheri and Jensen 2006). In short, single trials were convolved by a complex Morlet wavelet $w(t, f_0) = A\exp(-t^2/2\sigma_t)\exp(2i\pi f_0 t)$, where $\sigma_t = m/2\pi f_0$, and i is the imaginary unit. $A = (\sigma_t\sqrt{\pi})^{-1/2}$ is a normalisation term, whereas the constant m defines the time-frequency resolution tradeoff and was set to 7. The wavelet transformation produces a complex time series for the frequencies f_0 of interest (here 8-12 Hz). Single-trial power was calculated by averaging the squared absolute values of the convolutions in the 500 ms preceding the onset of the stimulus at the

subject-specific peak alpha frequency and occipitoparietal sensor with the highest overall alpha power.

Results

Our participants' behavioural performance indicated that our paradigm was successful in capturing choice confidence. Specifically, our participants selected the SR more frequently in more difficult trials ($F(2, 36) = 55.87$, $p < .001$, *post hoc* paired t-tests, all $p < .001$, Fig. 2.1B), consistent with previous reports showing that confidence scales with the amount of sensory evidence (Vickers and Packer 1982). Importantly, there was no difference in the frequency of choosing the SR across face and car trials ($t(18) = 1.7$, $p = 0.11$) ensuring this effect was not driven by one of the two stimulus categories.

More interestingly, accuracy on trials in which participants were offered the SR and rejected it was significantly higher compared to the trials in which the SR was not available ($F(1, 18) = 100.26$, $p < .001$, Fig. 2.1C). This effect was present for all levels of sensory evidence suggesting that participants waived the SR based on a sense of confidence on each trial rather than on the level of stimulus difficulty. Overall there was no significant difference in accuracy between face and car trials indicating that there was no category-specific choice bias ($t(18) = 0.76$, $p = 0.46$). As expected (Blank et al. 2013, Philiastides et al. 2006, Philiastides and Sajda 2006), there was also a main effect of stimulus difficulty ($F(2, 36) = 28.99$, $p < .001$, *post hoc* paired t-tests, all $p < .001$, Fig. 2.1C), with accuracy increasing with the amount of sensory evidence. Finally, we note, that due to the delayed-response paradigm employed here, there were no significant differences in response time between certain (SR Waived) and uncertain (SR Selected) trials (420ms and 406ms respectively, $t(18) = 0.99$, $p = .33$).

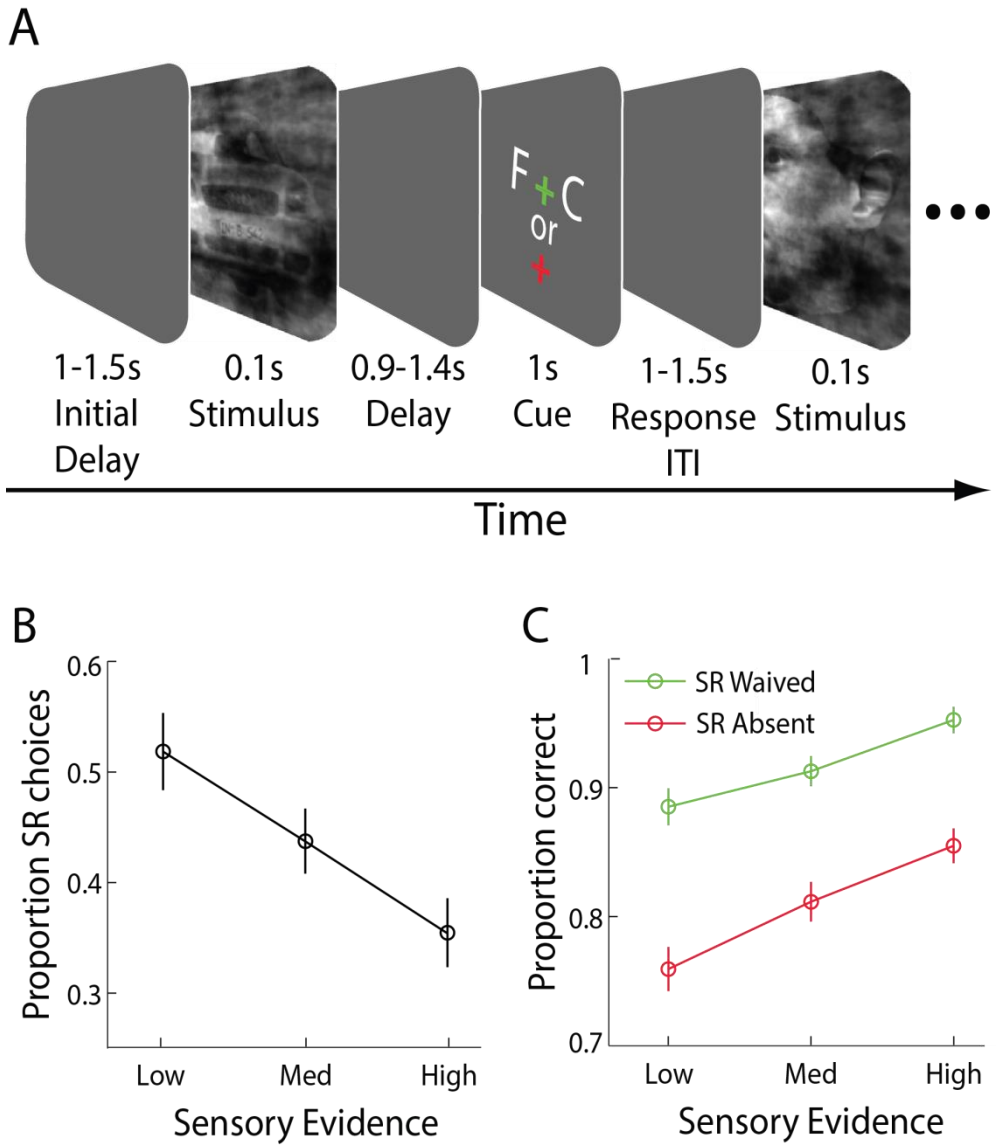

Figure 2.1. Experimental design and behavioural performance. **A.** Schematic representation of the behavioural paradigm. Participants had to categorise a briefly presented (0.1 s) image, at one of three possible levels of sensory evidence, as being a face or car. Stimulus presentation was followed by a random delay (0.9-1.4s) during which participants had to form a decision. Next, a visual response cue (1s) informed participants whether a small (relative to a correct choice) but sure reward (SR) was available or not, with either a green or red cross, respectively. The letters "F" (for face) and "C" (for car) where positioned randomly to the left and right of the fixation cross, indicating the mapping between stimulus and motor effectors (right index and ring fingers respectively). Participants indicated their choice as soon as the response cue was removed from the screen. **B.** Mean proportion of SR choices (on trials where the SR was offered), across subjects, as a function of sensory evidence. **C.** Mean proportion of correct responses, across subjects, for SR Waived (green) vs. SR Absent (red) trials, as a function of the three levels of sensory evidence. Error bars in B and C represent standard errors across subjects.

To identify confidence-related activity in the neural data, we used a single-trial multivariate approach to discriminate between certain (SR Waived) and uncertain (SR Selected) trials. We observed that the discriminator's performance increased gradually after 300 ms (i.e. after early encoding of the stimulus) and peaked around 600 ms post-stimulus, on average. This pattern of discriminator performance was visible in individual data (Fig. 2.2A) as well as in the group average (Fig. 2.2B), consistent with the idea that confidence develops gradually as the decision process unfolds and culminates before one commits to a choice (Ding and Gold 2013, Kiani and Shadlen 2009). To visualise the temporal profile of this discriminating component activity across trials, we also constructed single-trial component maps by applying our subject-specific spatial projections estimated in the time window yielding maximum confidence discrimination (using Eq. 1) to an extended time window. These maps clearly highlight the overall difference in component amplitude $\boldsymbol{y}$ between SR Waived and SR Selected trials and the temporally broad response profile of the discriminating component, both of which contributed to the discriminator's performance. The maps also highlight the trial-by-trial variability in the amplitude and temporal spread of this component, providing qualitative support that decision confidence might represent a graded quantity (Fig. 2.2C).

To provide further support linking this discriminating component to choice confidence, we considered trials in which the SR was not available (i.e. SR Absent) and participants were forced to make a face/car response. Importantly, these trials can be considered as "unseen" data (they are independent of those used to train the classifier), and can be subjected through the same neural generators (i.e. spatial projections) estimated during discrimination of SR Waived vs. SR Selected trials. We expected that these trials would contain a mixture of confidence levels and therefore the resulting mean component amplitude at the time of peak discrimination would be situated between those of the certain and uncertain trial groups (i.e. SR Waived > SR Absent > SR Selected). Indeed, this was the case and the mean SR Absent activity was significantly different from both the SR Selected ($t(18) = 7.53$, $p < .001$) and SR Waived ($t(18) = -7.71$, $p < .001$) (Fig. 2.2D). The mixture of both high and low

confidence trials within the SR Absent group can be further appreciated by inspecting the resulting single-trial component amplitudes (Fig. 2.2C; middle panel).

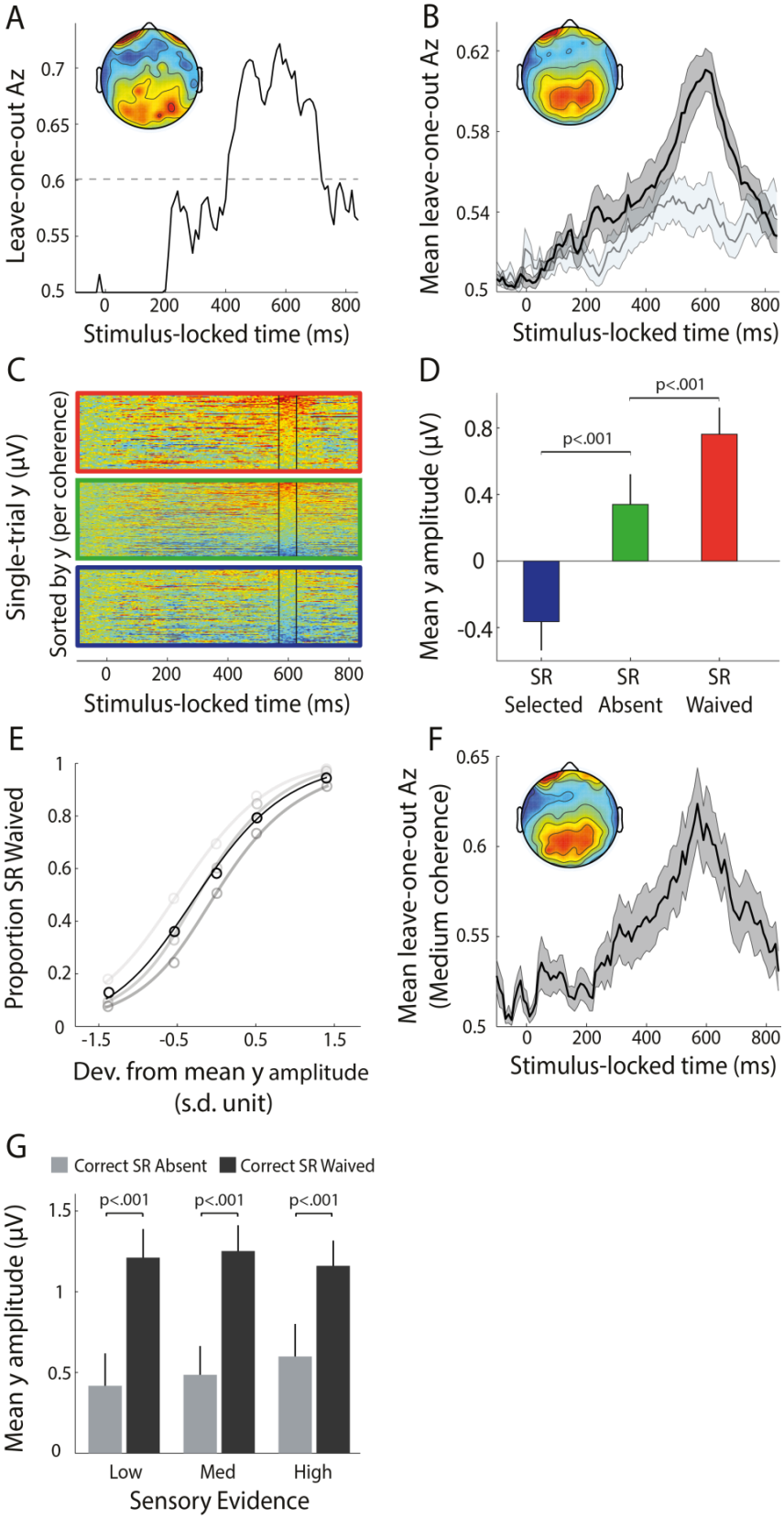

Figure 2.2. Neural representation of choice confidence. **A.** Classifier performance (Az) during high-vs-low confidence discrimination (i.e. SR Waived vs. SR Chosen) of stimulus-locked single-trial data, for a representative subject. The dotted line represents the subject-specific Az value leading to a significance level of p=0.01, estimated using a bootstrap test. The scalp topography is associated with the discriminating component estimated at time of maximum discrimination. **B.** Mean classifier performance and scalp topography across subjects during confidence (i.e. SR Waived vs. SR Chosen) discrimination (dark grey). For comparison, mean classifier performance during accuracy (i.e. Correct vs. Incorrect) discrimination for SR Absent trials is also shown (light grey). Shaded areas represent standard errors across subjects. **C.** Single-trial discriminant component maps, for a representative subject, obtained by applying the subject-specific spatial projections estimated at the time of maximum discrimination (black window) to an extended time range relative to the onset of the stimulus and across all trials (including SR Absent trials that were independent of those used to train the classifier). Each row in these maps represents discriminant component amplitudes, $y(t)$, for a single trial across time. Within each trial group (top to bottom panel: SR Waived, SR Absent, SR Selected), trials are sorted by mean component amplitude (y) at time of maximum discrimination. Red represents positive and blue negative component amplitudes, respectively. **D.** Mean component amplitude for the SR Absent group was situated between those of the high and low confidence groups (SR Waived and SR Selected). This is consistent with a mixture of "certain" and "uncertain" trials in the SR Absent group as can be seen in C for one participant (i.e. a mixture of red and blue component amplitudes). Error bars are standard errors across subjects. **E.** Trial-by-trial deviations from the mean component amplitude at time of maximum confidence discrimination were positively correlated with the probability of waiving the SR. To visualize this association the data points were computed by grouping trials into five bins based on the deviations in component amplitude. Importantly, the curve is a fit of Eq. 3 to *individual* trials. Grey curves are fits of Eq. 3 to each of the three levels of sensory evidence separately (light to dark grey represents high to low sensory evidence. **F.** Mean classifier performance and scalp topography across subjects within an individual level of sensory evidence (medium phase coherence; results looked very similar for the other two levels). Note that the patterns are qualitatively very similar to those shown in B for which classification was performed over all trials. Shaded area represents standard errors across subjects. **G.** Mean component amplitude for correct SR Waived (confident) trials (dark grey) and correct SR Absent (on average, less confident) trials (light grey), split by level of sensory evidence. Error bars are standard errors across subjects.

A potential concern is that subjects' choice to waive or select the SR (and consequently our discriminator's performance) is driven primarily by the physical properties of the stimulus (i.e. stimulus difficulty). This is unlikely, as changes in early stimulus encoding would have produced significant discrimination performance earlier in the trial (i.e. around 170–200 ms post-stimulus, driven by EEG components known to be affected by stimulus evidence – N170/P200

(Jeffreys 1996, Liu et al. 2000, Philiastides et al. 2006)), which was absent in our data (see discriminator performance at the relevant time windows in Fig. 2.2A, B). Nonetheless, we performed additional analyses to ensure that stimulus difficulty could not explain the observed effects.

We first removed the overall influence of stimulus difficulty by computing the trial-to-trial deviations around the mean discriminating component activity, separately for each level of sensory evidence, and used these residual fluctuations as predictors of participants' choices to waive the SR in a single-trial logistic regression analysis (Eq. 3). We found a significant positive correlation ($t(18) = 15.19$, $p < .001$) between component amplitudes and the probability of waiving the SR (i.e. bigger amplitudes, higher probability of SR waived; Fig. 2.2E). Crucially, we also repeated this regression analysis separately for each level of sensory evidence and found that our component amplitudes remained a significant predictor of subjects' opt-out behaviour within each level of stimulus difficulty (all $p < .001$), without significant differences in explanatory power across the three levels (all $p \geq .2$; Fig. 2.2E). Similarly, we repeated the discrimination between certain-vs-uncertain trials using observations from individual levels of sensory evidence and demonstrated that our discriminator performance remained virtually unchanged compared to our main analysis (compare Fig. 2.2B with 2.2F for a single level of difficulty).

To identify the spatial extent of our confidence component, we first computed a forward model of the discriminating activity (Eq. 2), which can be visualised in the form of a scalp map (Fig. 2.2A, B). Importantly, we used these forward models as a means of localizing the underlying neural generators using a Bayesian distributed source reconstruction technique (Friston et al. 2008). The source analysis revealed sources in areas in the anterior prefrontal cortex with a pronounced left bias and in regions of the posterior parietal cortex, bilaterally (Fig. 2.3; explained variance > 97%), areas which have previously been implicated in perceptual decision making and evidence accumulation, both in the human (Heekeren et al. 2006, Ploran et al. 2007, Tosoni et al. 2008) and primate (Kim and Shadlen 1999, Shadlen and Newsome 2001, Kiani and Shadlen 2009) brains. These results, coupled with the gradual build-up of confidence-

related discriminating activity (Fig. 2.2A, B), suggest that choice confidence might be encoded in the same brain areas supporting evidence accumulation and decision formation. Moreover, they raise the intriguing possibility that confidence is computed continuously as the decision process unfolds, thus being reflected in the slope of the process of evidence accumulation itself (Ding and Gold 2013, Kiani and Shadlen 2009).

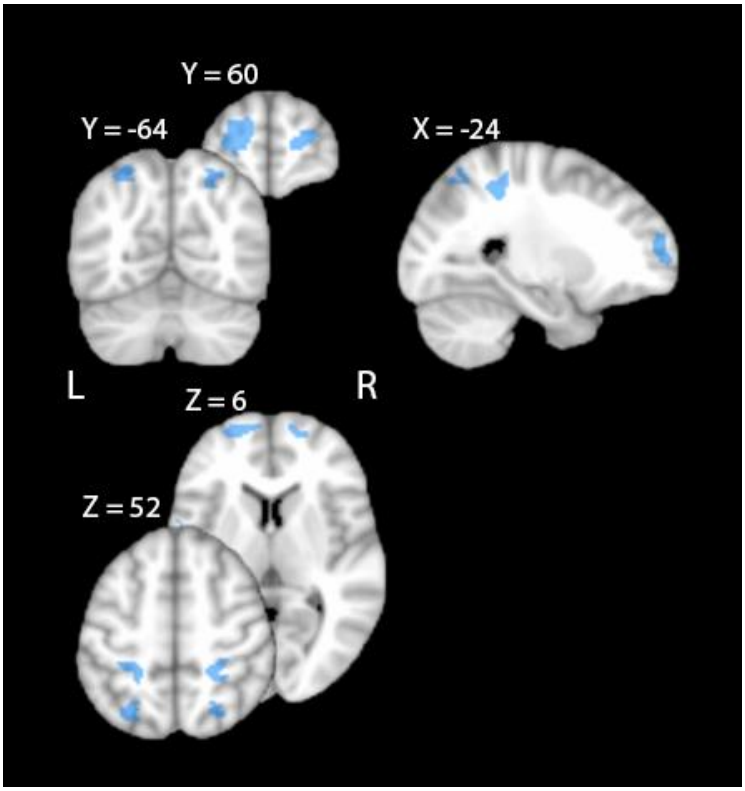

Figure 2.3. Spatial representation of choice confidence. A distributed source reconstruction technique (Friston et al. 2008) revealed neural generators associated with choice confidence in anterior prefrontal cortex (with a left bias) and in distinct clusters in parietal cortex, bilaterally (along the intraparietal sulcus). Slice coordinates are given in millimetres in MNI space.

To formally test these predictions, we subjected the data through the same neural generators (i.e. spatial projections) estimated for the confidence discrimination but stratified our trials along the sensory evidence dimension instead. In doing so, we observed ramp-like activity starting, on average, at 300 ms post-stimulus, which built up gradually to the time of peak confidence discrimination (Fig. 2.4A), and whose slope was parametrically modulated by the amount of sensory evidence ($F(2,36) = 10.6$, $p < 0.001$, Fig. 2.4B), consistent

with a process of evidence accumulation (Philiastides et al. 2006, Kelly and O'Connell 2013, Philiastides et al. 2014, O'Connell et al. 2012). Importantly, this finding suggests that choice confidence and evidence accumulation share common neural generators. To investigate whether confidence emerges from the decision process itself, we tested whether the trial-by-trial build-up rates of the accumulating activity were predictive of participants' opt-out behaviour. Specifically, we used single-trial slope estimates of the accumulating activity to predict participants' decisions to waive the SR in a new logistic regression model (Eq. 4). As in the previous analysis, overall stimulus difficulty effects were removed from individual trials. We found a significant positive correlation ($t(18) = 11.94$, $p < .001$) between the slope of accumulation and the probability of waiving the SR (i.e. steeper slopes, higher probability of SR waived, Fig. 2.4C).

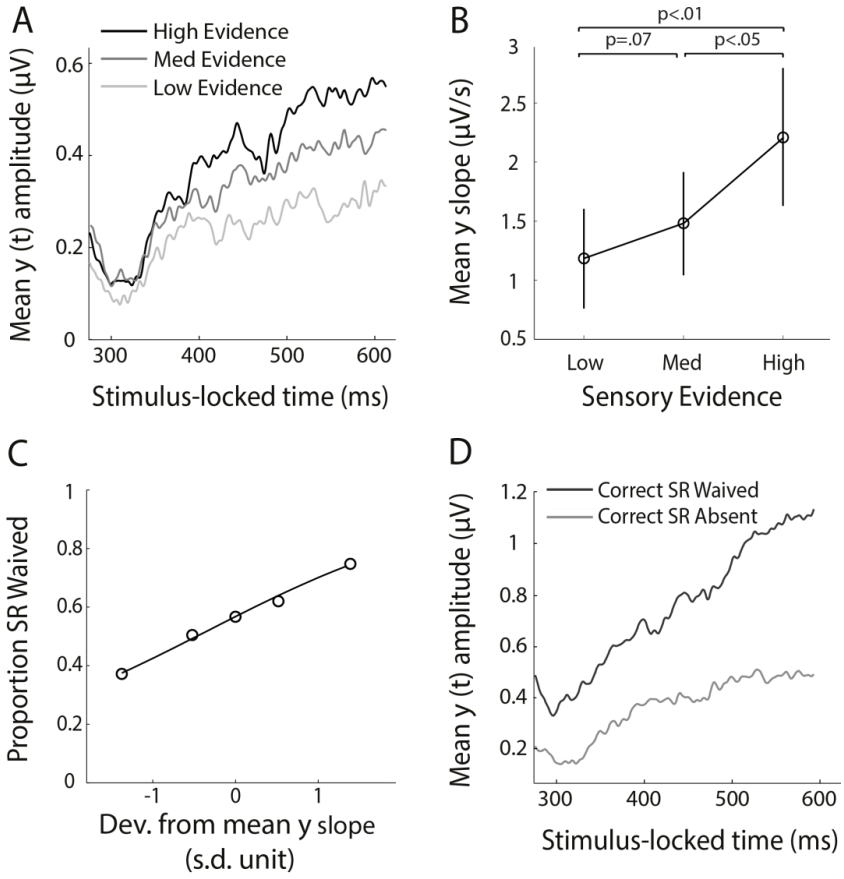

Figure 2.4. Choice confidence and evidence accumulation. **A.** Subjecting our data through the same spatial distribution of component activity estimated during confidence discrimination (i.e., Fig. 2.2A, B) revealed a gradual build-up of activity (i.e. accumulating activity) earlier in the trial that was modulated by the amount of sensory

evidence (i.e. % stimulus phase coherence). Trials were locked to the onset of the stimulus and averaged across subjects. **B.** Mean slope of the accumulating activity across subjects was positively correlated with the amount of sensory evidence. Slopes were estimated by computing linear fits through the data based on subject-specific onset and peak accumulation times. Error bars represent standard errors across subjects. **C.** Trial-by-trial deviations from the mean accumulation slope were positively correlated with the probability of waiving the SR. To visualize this association the data points were computed by grouping trials into five bins based on the deviations in the slope of the accumulating activity. Importantly, the curve is a fit of Eq. 4 to *individual* trials.

A potential confound of the previous analysis is that the slope of the accumulating activity simply echoes the confidence effects we identified earlier on the amplitude of our discriminating component, as the latter were extracted, on average, near the end of the accumulating activity. Crucially, we found that the two quantities were only partially correlated ($r = .39$, $p < .001$), due to the high degree of inter-trial variability in internal components of decision processing as has been described previously by accumulation-to-bound models (Ratcliff et al. 2009, Bogacz et al. 2006, Mulder et al. 2014, van Maanen et al. 2011). As such we found that each exerted a separate influence on our participants' opt-out behaviour (Eq. 5, $t (18) = 2.96$, $p = .008$), which suggests that traces of confidence begin to develop as early as the decision process itself and continue to be reflected in the process of evidence accumulation, becoming progressively more robust as the decision unfolds.

Importantly, to rule out that our confidence effects are driven by changes in attention across individual trials we exploited two well-known neural signatures of attention (pre-stimulus alpha (Wyart and Tallon-Baudry 2009) and a post-stimulus evoked response indexing allocation of attentional resources (Philiastides et al. 2006)), which we used as additional predictors of our participants' opt-out behaviour in a different logistic regression model (Eq. 6). Crucially, we found that our original confidence component amplitudes and accumulation slopes remained significant predictors of the likelihood of waiving the SR (component amplitudes: $t (18) = 14.51$, $p < .001$, one-tailed t-test; slopes: $t (18) = 2.15$, $p < .05$). Furthermore, to test whether local fluctuations in attention could further explain our findings, we used a serial autocorrelation

regression analysis to predict our discriminator component amplitudes ($\boldsymbol{y}$) on the current trial using those on the immediately preceding five trials and found no significant effects (all $p > 0.1$). Taken together, these results provide compelling evidence that our observed effects could not be explained purely by changes in attention.

To ensure that accuracy, which was shown to correlate partially with decision confidence (Vickers et al. 1985, Vickers and Packer 1982), is not responsible for the reported effects, we performed two additional control analyses. First, we used SR Absent trials, which contained trial-to-trial accuracy information and trained a separate classifier to discriminate between correct and incorrect trials. If our confidence effects were a mere manifestation of differences between correct and incorrect trials then classification performance would have been comparable to that obtained along the confidence dimension. Instead, classifier performance was significantly reduced relative to our SR Waived vs. SR Selected discrimination (Fig. 2.2B, $t(18) = 5.1$, $p < .001$, paired t-test).

Finally, we performed an analysis in which we subjected the data through the same neural generators (i.e. spatial projections) estimated for the confidence discrimination and partitioned our trials in two groups in a way that ensured accuracy remained constant while confidence was altered across the groups. Specifically, we compared component activity between correct SR Waived trials (confident trials) and correct SR Absent trials (which are, on average, less confident as they contain a mixture of confident and non-confident choices). We found that the component amplitudes for the more confident group of trials were significantly higher ($t(18) = 9.4$, $p < .001$, paired t-test) with persistent effects across all levels of sensory evidence (Fig. 2.2G, *post hoc* paired t-tests, all $p <. 001$). Taken together, these results endorse the notion that our reported confidence effects cannot be explained purely by differences in decision accuracy.

Discussion

Here, we used a multivariate single-trial EEG approach, coupled with a distributed source reconstruction technique, to provide a mechanistic account on how decision confidence is represented in the human brain. We showed that a neural representation of confidence arises as early as the decision process itself and becomes progressively more robust as the decision unfolds, culminating shortly before one commits to a choice. Importantly, we demonstrated that this representation is reflected in the rate of evidence accumulation, thereby linking the development of choice confidence to the same neural mechanism used to form the decision itself. Consistent with this interpretation, source reconstruction placed confidence-related activity in regions previously implicated in evidence accumulation and decision making in human prefrontal and parietal cortices (Heekeren et al. 2006, Ploran et al. 2007, Filimon et al. 2013, Tosoni et al. 2008).

Together, these findings lend support to the idea that there exists a general-purpose decision making network involved in accumulating evidence for a decision while simultaneously encoding the confidence in that decision. Overall, our findings are in line with a recent report showing that neurons in lateral intraparietal cortex of the primate brain represents the formation of the decision as well as the degree of confidence underlying that decision (Kiani and Shadlen 2009). Similarly, a growing body of evidence from animal neurophysiology suggest that when the brain forms a decision it does so in a way that resembles a Bayesian inference, in the sense that even for binary choices, a decision is formed by sampling and gradually accruing information from probability distributions rather than single estimates representing each of the alternatives (Ma et al. 2006, Zemel et al. 1998). In this framework, a measure of confidence arising directly from the decision process itself can therefore be thought of as a graded quantity, representing degree of belief that an impending choice will be correct.

Key to establishing a quantitative association between decision confidence and neural activity was our ability to exploit single-trial information within each

class of nominally identical stimuli, thereby controlling for confounding effects of stimulus difficulty and attention. Specifically, we demonstrated that trial-by-trial fluctuations in confidence-related neural activity remained predictive of opt-out behaviour even after accounting for the overall amount of task difficulty as well as when extracted and tested separately for each level of sensory evidence. Similarly, we addressed the possibility that our confidence effects merely reflected changes in participants' attentional state on each trial, either prior to, or during the decision process.

In doing so, we considered two neural measures, which have previously been hypothesized to reflect top-down influences of attention on the decision process during visual discriminations, and investigated the extent to which they predicted participants' choice confidence (i.e., opt-out behaviour). Importantly, we showed that neither of these measures hindered the explanatory power of the confidence discriminating neural activity. Likewise, we also showed that local fluctuations in attention *across* trials, as assessed via a serial autocorrelation regression analysis, could not provide an adequate account of our findings. Whilst we do not dismiss the possibility that trial to trial variability in attention may exert a top-down influence on the efficiency of stimulus encoding and/or decision process, and ultimately on the level of confidence in one's choice, our findings render a purely attentional account of the observed confidence effects unlikely.

Although we designed our experiment to discourage explicit updating of reward expectations (i.e., we did not provide feedback as to whether a choice was correct or not) it remains possible that our representation of choice confidence can be explained by the expected value of the chosen option in so far as the latter is correlated with one's belief that their choice is correct. In fact, it has recently been suggested that structures such as the ventromedial prefrontal cortex may use a common neural currency to represent both confidence and value associated with a choice (Lebreton et al. 2015). Nevertheless, the regions we identified here as being associated with choice confidence (using source reconstruction analysis) appear to be located outside the networks most

commonly associated with expected reward and value signals (Dreher et al. 2006, Kable and Glimcher 2007, Knutson et al. 2005, Rangel et al. 2008, Rolls et al. 2008, Rushworth and Behrens 2008, Philiastides et al. 2010).

Finally, our findings that confidence signals appear as early as the process of evidence accumulation itself constitute evidence against a purely metacognitive (post-decisional) account of decision confidence, and are consistent with a recent study showing temporal overlap of confidence- and decision-related electrophysiological signal during perceptual decisions (Zizlsperger et al. 2014). Importantly, however, our results do not exclude the possibility that confidence representations persist beyond the decision point and after a behavioural choice was made (Fleming et al. 2012, Pleskac and Busemeyer 2010). Nonetheless, these metacognitive representations are captured using post-decisional subjective confidence reports, which are likely to be subjected to additional influences arriving after the decision point (e.g. internal noise, expected reward etc.). In addition, the extent to which these post-decisional signals influence metacognitive assessment and subsequent choices remains unclear. Future studies designed to investigate how decisional and post-decisional confidence signals interact to shape behaviour would be necessary. In particular, understanding how confidence traces arising from the process of decision formation are communicated to regions implicated in metacognitive appraisal would be required (Fleming et al. 2012, De Martino et al. 2013, Hebart et al. 2014).

In summary, choice confidence represents the degree of belief that one's actions are likely to be correct and as such can play a critical role in how we interact with the world around us. Here, we provided a mechanistic account on how confidence is represented in the human brain and provided strong evidence that linked the development of choice confidence to the same mechanism and neural generators used to form the decision itself. These results could provide the foundation upon which future computational studies could continue to interrogate the mechanistic details of the influence of confidence on decision making (Zylberberg et al. 2012). Crucially, our findings coupled with our ability

to exploit the relevant neural signatures non-invasively and on a trial-by-trial basis, may have direct implications for decision-making problems that rely on inconclusive or partially ambiguous evidence. Specifically, they can provide the platform for developing cognitive interfaces that can help facilitate, and ultimately optimise decision making (Sajda et al. 2009, Sajda et al. 2007).

Chapter 3. Alpha- and beta-band oscillatory activity reflects neural representations of confidence in perceptual decisions

Summary

Confidence in a perceptual decision represents an internal estimate of accuracy, and as such can play an essential role in informing relevant goal-directed actions. Electrophysiological studies have shown that oscillatory patterns in the neural activity that characterises perceptual decisions contains valuable information about its underlying neural mechanisms, however it is not clear whether these rhythmic fluctuations may also be informative about the associated confidence. The current study adopted an exploratory approach to address this question. We used a post-decision wagering paradigm to behaviourally separate high- from low-confidence choices. Specifically, subjects made face vs. car visual categorisations and were rewarded for correct responses, with the possibility to opt out of the task for a smaller but certain reward on a random subset of trials. Subject's decision to use or refuse this option indicated confidence (low and high, respectively) in their judgment. Importantly, the perceptual decision and motor response stages were separated by a forced delay during which the exact response mapping remained unknown. We identified confidence-discriminating oscillatory activity in the alpha and beta bands. This was most prominent over the sensorimotor electrodes contralateral to the motor effector (i.e., right hand) used for indicating choice. The effect was transient in nature, peaking before subjects could plan a response, and appeared to overlap in time with a (separate) confidence-related signal which we previously identified in the time-domain, and which was shown to reflect both the decision formation and associated confidence. Together, these results open the possibility that motor systems may track both the evolving perceptual decision and formation of confidence, in preparation for impending action.

Introduction

Confidence in a perceptual decision represents an individual's subjective assessment of the likelihood that a judgment is correct, and as such can contribute significantly to guiding relevant actions. A driver may, for instance, initiate a press of the brake pedal based on the strength of belief that the object spotted in the distance on the road is an animal and not a shadow.

The electrophysiological correlates of perceptual confidence have been investigated in both human and animal subjects, showing that confidence-relevant information can be identified early in the decision process, and as early as the decision process itself (Kiani and Shadlen 2009, Gherman and Philiastides 2015, Zizlsperger et al. 2014, Middlebrooks and Sommer 2012). In the previous chapter we identified neural activity in the EEG signal which reliably discriminated between high- and low-confidence perceptual decisions. This activity exhibited a ramp-like temporal profile modulated by the strength of sensory evidence, thus resembling a gradual build-up of decision-related evidence accumulation, as observed by previous EEG studies (O'Connell et al. 2012, Twomey et al. 2015, Kelly and O'Connell 2013, Philiastides et al. 2014), and consistent with the idea that confidence relies on information contained in the decision process itself (Kiani and Shadlen 2009).

While time-domain analyses of neural activity can offer valuable insights into the temporal properties and underlying mechanisms of confidence-related processes, the rhythmic properties inherent to these signals can potentially contain complementary information. Spectral analysis can be better suited for characterising sustained modulations of relevant oscillatory signals which are not phased-locked to, but still induced by, perceptual events of interest (Donner and Siegel 2011, Pfurtscheller and Lopes da Silva 1999). Spectral markers of perceptual decision processes have been identified across different frequency ranges, including theta (Van Vugt et al. 2012), beta (Haegens et al. 2011, Donner et al. 2009), and gamma bands (Donner et al. 2009, Polania et al. 2014). In particular, these studies have revealed that the magnitude of oscillatory activity scales with both stimulus properties and choice-related behaviour (i.e., decision

time and performance), in a fashion that resembles an accumulation of decision-related evidence. Of particular interest are several studies which have shown that oscillatory activity in the alpha and beta frequencies, recorded over motor and premotor regions, can exhibit similar characteristics in anticipation of an impending response (Donner et al. 2009, Wyart et al. 2012, O'Connell et al. 2012, Haegens et al. 2011, de Lange et al. 2013). These effects have been observed over the hemisphere contralateral to the motor effector, and consist in the suppression in the oscillatory power during perceptual stimulation and decision formation stages, with the magnitude of the suppression scaling with the strength of sensory evidence, and behavioural performance. Together, these observations have led to the view that the motor system may track the decision process as it forms (i.e., in a continuous fashion), in support for an impending action (Gold and Shadlen 2007, Gold and Shadlen 2000, Siegel et al. 2011).

As the rhythmic dynamics of perceptual decisions are being uncovered, the question emerges whether these oscillatory signals may also carry information about the confidence associated with the decision. Indeed, taking together the observation that decision-related neural signals hold information about eventual confidence (see Chapter 2), and the finding that motor systems carry information about the decision process, one might predict that confidence-related information may also be encoded in the motor system. Currently, the spectral dynamics of perceptual confidence are not well understood. Confidence-discriminating activity has been detected in the gamma band using intracranial EEG (Peters et al. 2017). In addition, there is evidence that oscillatory activity in the theta band may support metacognitive processes (Wokke et al. 2017), with midfrontal theta signalling task performance via increases in power after detected errors (Cohen 2016, Murphy et al. 2015). Finally, a recent study (Kubanek et al. 2015) has demonstrated that confidence in an auditory decision can be inferred from motor-selective alpha-band activity before a response is initiated, suggesting information in lower frequency bands may encode not only decision-related information, but also confidence in the impending action.

Here, we aimed to investigate whether confidence-discriminating oscillatory activity can be detected during visual decision making, and if so, how this might relate to confidence-related evoked responses in the time-domain. We performed analyses on an existing dataset (results from the previous study are reported in Chapter 2). Subjects performed a delayed-response post-decision wagering task (Kiani and Shadlen 2009, Fetsch et al. 2014) whereby they made perceptual categorisations of face-vs.-car stimuli and were rewarded for correct choices. On a random subset of trials, subjects were allowed to withhold from making a stimulus choice by selecting a smaller but certain reward instead. Behaviour on these trials was used as an indicator of confidence, with subjects' decision to exercise or ignore the opt-out choice reflecting low or high confidence in the perceptual judgment, respectively. Using a frequency-analysis approach, we identified confidence-discriminating activity over electrode sites corresponding to the motor effector used subsequently to express choice. These signals were transient in nature and were only observed during the decision stage of the trial (i.e., considerably in advance of an overt behavioural response). These were independent of the strength of stimulus evidence or spontaneous fluctuations in the prestimulus period. Interestingly, the timing of these signals appeared to coincide with a non-oscillatory confidence-related signal identified previously over parietal electrodes. Our results thus suggest that motor systems may carry information about both the evolving perceptual decision and associated confidence.

Materials and Methods

This study is based on reanalysis of data presented in Chapter 2 (see also Gherman and Philiastides, 2015). All methodological details relating to participants, stimuli and behavioural paradigm, as well as EEG data acquisition and pre-processing, are identical unless otherwise specified.

Participants. Nineteen healthy paid volunteers (7 males, mean age = 23.4 years, range 18-36) participated in the study. All had normal or corrected-to-normal vision and reported no history of neurological problems. Informed consent was

obtained from all subjects in accordance with the School of Psychology Ethics Committee at the University of Nottingham.

Stimuli and task. Stimuli and the behavioural paradigm are described in more detail in Chapter 2 (Fig. 2.1A). In short, the behavioural task dissociated between high- and low-confidence perceptual judgments by means of a post-decision wagering method (Kiani and Shadlen 2009). Specifically, subjects made delayed-response visual categorisations of noisy face and car stimuli, and received rewards in the form of points (correct response = 10 points; incorrect response = 0 points), which were converted into monetary bonuses at the end of the experiment. Importantly, on a random half of the trials, an additional response option was available, which allowed subjects to opt out of the face/car choice in exchange for a smaller (8 points) but certain reward (henceforth referred to as the sure reward, SR). The goal of this manipulation was to encourage participants to select the sure reward option on these trials if they were uncertain of their perceptual judgment, and provide a face/car response otherwise. Importantly, subjects were not aware in advance whether this option would be available.

On each trial, the visual stimulus was presented for 100 ms, followed by a random forced delay of 900-1400 ms (i.e., the decision phase). Next, a visual cue (1000 ms) informed subjects of their response options. Specifically, a red-coloured central fixation cross indicated that the SR option was not available and thus a face/car choice response was required, whereas a green-coloured cross indicated that the SR option was available. Letters “F” (face) and “C” (car) located randomly on the left and right sides of the fixation cross informed subjects of the mapping between stimulus and motor effector (index and ring fingers, respectively). This served to reduce potential confounds related to motor preparation processes during the decision phase of the trial. A response was only permitted once the response cue disappeared, during the inter-trial interval (1000-1500 ms), and thus at least 2000 ms after onset of the stimulus. Subjects made all responses using one of three buttons on a button box, namely a left/right press for a providing a stimulus response, and a central press for exercising the SR option.

EEG data acquisition. We recorded EEG data during performance of the main task, in an electrostatically shielded room, using a DBPA-1 digital amplifier (Sensorium Inc., VT, USA), at a sampling rate of 1000Hz. We used 117 Ag/AgCl scalp electrodes and three periocular electrodes placed below the left eye and at the left and right outer canthi. Additionally, a chin electrode was used as ground. All channels were referenced to the left mastoid. Input impedance was adjusted to <50kOhm. To obtain accurate event onset times we placed a photodiode on the monitor to detect the onset of the stimuli. An external response device was used to collect response times. Both signals were collected on two external channels on the EEG amplifiers to ensure synchronization with the EEG data.

EEG data pre-processing. We applied a 0.5-100Hz band-pass filter to the data to remove slow DC drifts and high frequency noise. These filters were applied noncausally (usingMATLAB "filtfilt") to avoid phase related distortions. Additionally, we re-referenced our data to the average of all electrodes. To remove eye movement artefacts, participants performed an eye movement calibration task prior to the main experiment, during which they were instructed to blink repeatedly several times while a central fixation cross was displayed in the centre of the computer screen, and to make lateral and vertical saccades according to the position of the fixation cross. We recorded the timing of these visual cues and used principal component analysis to identify linear components associated with blinks and saccades, which were then removed from the EEG data (Parra et al. 2005).

EEG spectral analysis. Spectral analyses were performed using the FieldTrip toolbox (Oostenveld et al. 2011), and custom MATLAB (MathWorks) code. Pre-processed data were segmented into epochs from -1000 to 1500 ms relative to the onset of the face/car stimulus. We computed the time–frequency representations of the EEG signal at 49 frequencies (4-100 Hz, in steps of 2 Hz), using a sliding-window Fourier transform. Time-frequency decomposition was performed separately for each channel and trial, on time windows centred from

-300 ms to 1000 ms (step-size of 50 ms) relative to the onset of the face/car stimulus (i.e., the decision stage). Prior to the Fourier transform, windows of interest were multiplied with a single Hanning taper. In an effort to maintain an optimal balance between spectral and temporal resolution of the time-frequency power estimates (Cohen 2014), the length of the sliding window was adapted to each frequency. Specifically, we used logarithmically-spaced numbers of cycles (rounded to the nearest integer), ranging from 4 cycles (1000 ms) for the lowest frequency (4Hz), to 16 cycles (160 ms) for the highest frequency (100Hz).

We also performed a separate control analysis where we further optimised parameters for detecting relevant activity in the high frequency ranges (30-100 Hz). Namely, we computed time-frequency representations using the multitaper approach (Mitra and Pesaran 1999), with three orthogonal slepian tapers, a sliding fixed window length of 250 ms, and frequency smoothing of ±8Hz. The multitaper method can be better suited for estimating frequency representations characterised by low signal-to-noise ratio, as is the case with oscillatory signals in the higher frequency range (Cohen 2014). Nonetheless, this adjustment did not change our results (i.e., no effects of confidence were observed in the 30-100 Hz range).

Single-trial power estimates resulting from the time-frequency decomposition were averaged across trials, separately for each trial group of interest (i.e., SR_{WAIVED} waived vs. $SR_{SELECTED}$) and subject. Resulting values were subsequently baseline-normalised using a decibel (dB) transform: dB=10*log10(Power/Baseline). The baseline was defined as the average power estimated from 5 sliding windows (step-size of 50 ms) in the -500 to -300 ms interval prior to stimulus onset, and using the same frequency-specific window lengths as for the post-stimulus period. Baseline normalisation was performed individually for each channel and frequency. To increase signal-to-noise ratio of the baseline estimate, this was computed by averaging across all experimental trials. Thus, normalisation was performed in a non-condition-specific manner, preserving potential power differences between the trial groups of interest in the prestimulus interval. Baseline-normalised condition averages obtained this

way, in the form of channel-frequency-time sets, were used for further statistical analysis.

Statistical analysis. We compared oscillatory power in the two trial groups of interest, namely SR_{WAIVED} and $SR_{SELECTED}$, and used a non-parametric cluster-based permutation approach (Maris and Oostenveld 2007) as implemented by the FieldTrip toolbox, to assess significance of the results. This method aims to address the multiple comparisons problem resulting from testing for effects at multiple electrode sites, frequencies and time points, with increased statistical power compared to more conservative procedures such as the Bonferroni correction. It does so by clustering data samples exhibiting similar effects according to their adjacency in space, time and/or frequency. For every channel-frequency-time sample, SR_{WAIVED} and $SR_{SELECTED}$ averages were compared across subjects by means of a paired t-test. All samples whose t-values exceeded a cluster-defining threshold of $\alpha_{THRESHOLD}$=.001 (two-sided tests) were grouped into channel-frequency-time clusters (minimum of 2 channels per cluster). The cluster-level summary statistic was defined as the sum of all t-values within each cluster obtained this way.

Finally, significance was established by comparing the cluster-level summary statistic against the randomisation null distribution resulting from 2000 random permutations. Namely, for each iteration, SR_{WAIVED} and $SR_{SELECTED}$ averages were permuted within each subject, and the maximum cluster-level summary statistic was used to build the randomisation null distribution. Clusters in the observed data which exceed the family-wise error-corrected threshold of $\alpha_{CLUSTER}$=.01 were considered significant.

Results

Behaviour. All behavioural results are presented in the Results section of Chapter 2. Importantly, we showed that responses on the decision task were more likely to be correct on trials where subjects were offered the SR option and declined it (i.e., by indicating a stimulus choice), compared to trials in which the SR option was not available to begin with ($F(1, 18) = 100.26$, $p < .001$,

see Fig. 2.1C). This finding is consistent with confident choices being more accurate (Baranski and Petrusic 1998), and thus indicate that the paradigm accurately captured subjects' confidence in their perceptual decisions. Additionally, we showed that when sensory evidence was weaker, subjects selected the SR more often ($F(2, 36)=55.87$, $p<.001$, post hoc paired t-tests, all p b .001, see Fig. 2.1B), suggestive of decreasing confidence (Vickers and Packer 1982, Festinger 1943).

Frequency analysis. The cluster-based permutation analysis revealed two channel-time-frequency clusters (Fig. 3.1) where oscillatory activity differed significantly between the SR_{WAIVED} vs. $SR_{SELECTED}$ trial groups ($p_{CLUSTER}=.001$ and $p_{CLUSTER}=.002$, respectively). Together, these spanned the alpha and beta (~10-22 Hz) frequency bands and showed a negative effect of confidence, i.e., a reduction of oscillatory activity in high- compared to low-confidence trials (Fig. 3.1). For the first cluster, this difference was most pronounced in the alpha and low-beta frequency ranges (~10-16 Hz) between approximately 550-650 ms relative to stimulus onset (Fig. 3.1, white dots), and located over parietal and left centro-posterior electrodes. For the second cluster, the difference was stronger in the mid-beta frequency range (~18-22 Hz) between ~500-600 ms post-stimulus, and appeared more localised in the left centro-lateral sensors (Fig. 3.1, orange dots).

We performed additional control analyses by focusing on two subsets of the spatio-temporo-spectral data that showed confidence-discriminating activity, as informed by our cluster-based permutation analysis. Specifically, we extracted the data (i.e., power estimates) from all channels, frequency bins, and time windows covered by each of the two clusters (see Fig. 3.1). For the first cluster (henceforth referred to as Cluster 1), data were extracted from parietal and left centro-posterior electrodes (Fig. 3.2, top electrode map), within the 10-16 Hz (i.e., alpha/beta) frequency range, and the 550-650 ms time window, whereas for the second cluster (Cluster 2), we focused on data from left centro-parietal electrodes (Fig. 3.2, bottom electrode map), within the 18-22 Hz frequency range, and the 500-600 ms time window.

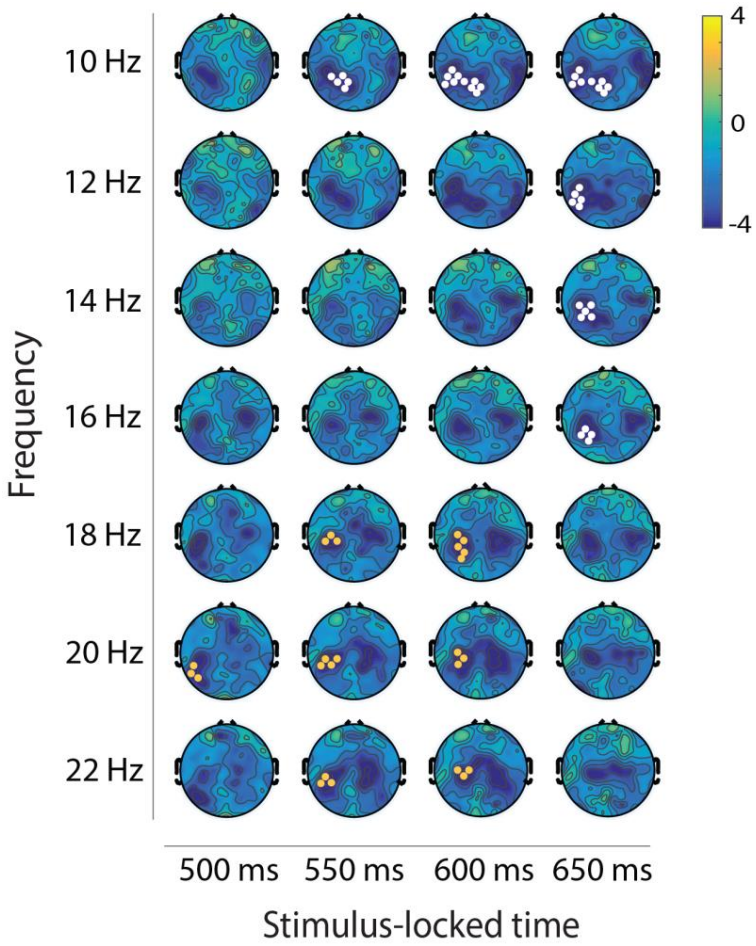

Figure 3.1. Confidence-discriminating spatio-temporo-spectral clusters obtained with the cluster-based permutation analysis ($p_{CLUSTER}$<.01). Colours represent average t-values resulting from the subject-level paired comparisons.

To visualise the time-frequency representation of the confidence-discriminating activity, we averaged our data subsets across the spatial (channel) dimension. Results are displayed in Fig. 3.2A. Next, we evaluated the temporal profile of the confidence-discriminating activity relative to baseline, by averaging data across both the spatial and spectral dimensions. As can be observed in Fig. 3.2A and Fig. 3.2B, a suppression of oscillatory activity was present in both trial groups following stimulus onset, which was more pronounced for SR_{WAIVED} choices. To formally test this effect whilst avoiding circularity, we performed individual comparisons of the SR_{WAIVED} and $SR_{SELECTED}$ trial groups, with a separate set of trials which were not used in the cluster analysis (i.e., "unseen" data), namely those in the SR_{ABSENT} condition where subjects were not offered the possibility to opt out of the decision. As these trials are likely to contain a

mixture of certain and uncertain choices, we therefore expected corresponding oscillatory power to be situated, on average, between that of the SR_{WAIVED} and $SR_{SELECTED}$ conditions. We performed group-level paired t-tests at each time point within the decision period, and found significant differences ($p<.05$, uncorrected, see Fig. 3.2B) across extended time windows, thus suggesting that confidence-discriminating activity observed here is unlikely to be merely artifactual.

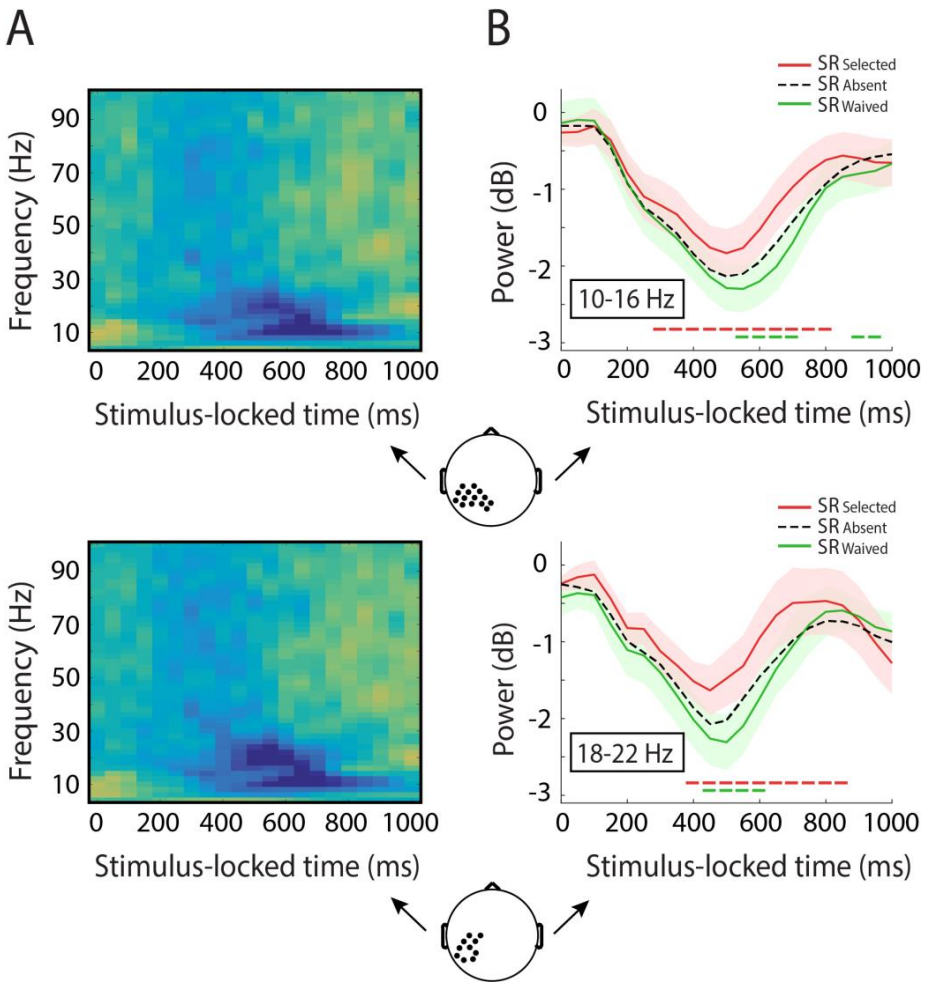

Figure 3.2. Confidence-discriminating oscillatory activity. Top and bottom figures represent data subsets extracted on the basis of the two clusters identified in the cluster-based permutation analysis, respectively. **A.** Time-frequency representation of baseline-normalised power. Colours represent t-values resulting from the subject-level paired comparisons. **B.** Time course of oscillatory power, separated by trial groups of interest. Markers running along the bottom of the plot represent significant ($p<.05$, uncorrected) differences between the SR_{ABSENT} and SR_{WAIVED}/$SR_{SELECTED}$ trials (green and red, respectively), for each time point of interest.

Overall, the spatio-temporo-spectral pattern of the observed oscillatory activity is qualitatively similar to the well-established phenomenon known as event-related desynchronisation (Pfurtscheller and Lopes da Silva 1999), which refers to the suppression of oscillatory power in the upper alpha (~8-14 Hz; also known as "mu") and beta (~15-30 Hz) bands, associated with motor processing. This effect, which is observed over sensorimotor regions and typically lateralised to the hemisphere contralateral to the motor effector, is hypothesised to play a role in the representation, preparation, and execution of movement (Cheyne 2013, Pfurtscheller and Lopes da Silva 1999, McFarland et al. 2000, Alegre et al. 2003, Neuper and Pfurtscheller 2001), though interestingly a few studies have also shown that its evolution in time can reflect the formation of perceptual decisions that inform associated actions (Donner et al. 2009, O'Connell et al. 2012, de Lange et al. 2013).

Confidence-related desynchronisation effects within both clusters appeared transient in nature (see Fig. 3.2B), with oscillatory power for SR_{WAIVED} and $SR_{SELECTED}$ converging to the same near-baseline values before 1000 ms post-stimulus, and thus prior to the end of the decision phase and onset of the response-informative cue (note that the response-informative cue was presented at least 1000 ms after stimulus onset). Additionally, there was no evidence of confidence-related effects in the time period leading to a behavioural response, as assessed with a separate cluster-based permutation analysis in which data were locked to the onset of response. These observations remained true even at considerably more liberal thresholds ($\alpha_{THRESHOLD}$=.05, two-sided test; $\alpha_{CLUSTER}$=.05). While the latter is likely a consequence of the forced delay employed in this paradigm, and therefore unsurprising, overall the above observations suggest that the confidence-related suppression in oscillatory power is unlikely to be linked to subjects' overt motor responses during the response stage of the trial.

Strength of sensory evidence. Motor-preparatory activity in the alpha and beta frequency bands has previously been shown to be modulated by the strength of sensory evidence that informs subsequent choice (de Lange et al. 2013). To test whether stimulus difficulty alone may explain our confidence-discriminating oscillatory activity, we removed this influence in our data by extracting trial-to-trial fluctuations around the mean power estimates (i.e., z-scores) within each level of sensory evidence. We then repeated the cluster-based permutation analysis on the resulting values, as detailed above. We found that our results remained both qualitatively and quantitatively very similar (Fig. 3.3), with the two clusters continuing to show significant differences between SR_{WAIVED} and $SR_{SELECTED}$ trial groups ($p_{CLUSTER}$=.002 for both clusters). This suggests

that the observed effect cannot be purely explained by the physical properties of the stimulus.

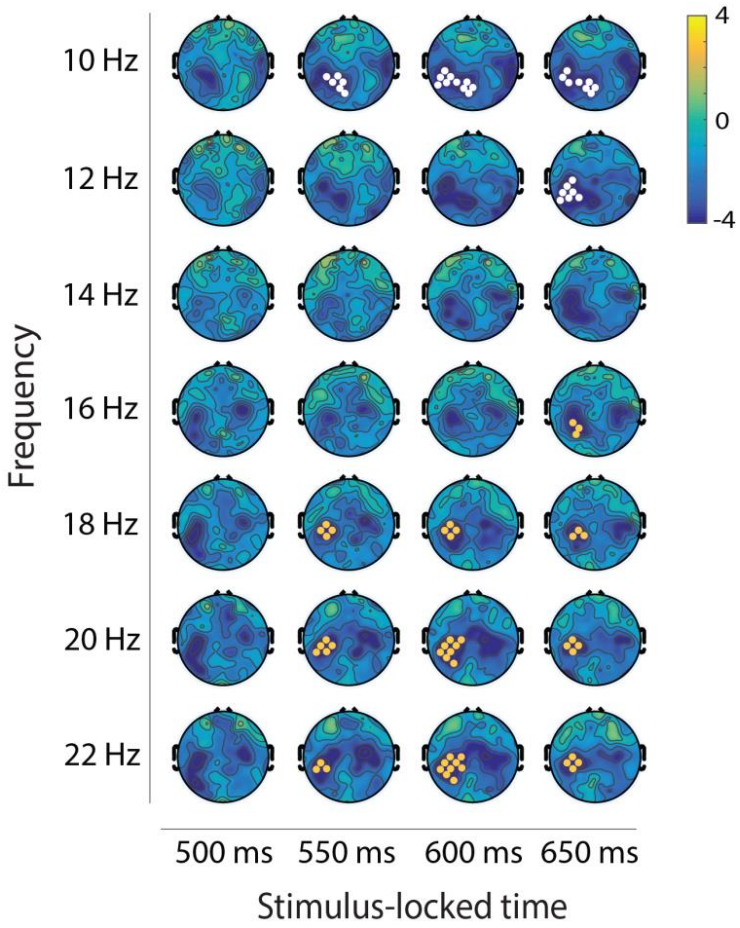

Figure 3.3. Confidence-discriminating spatio-temporo-spectral clusters obtained with the cluster-based permutation analysis ($p_{CLUSTER}$<.01), where influences of task difficulty have been removed (see 'Strength of sensory evidence' subsection in Results). Colours represent t-values resulting from the subject-level paired comparisons.

Prestimulus states. Alpha- and beta-band desynchronisation over motor/premotor regions can occur spontaneously (i.e., prior to stimulus presentation), affecting both the oscillatory activity during the perceptual decision, and associated behaviour (de Lange et al. 2013). Additionally, prestimulus fluctuations in the alpha-band have also been shown to affect confidence in upcoming perceptual decisions (Baumgarten et al. 2016, Samaha et al. 2017). To test whether our results are independent of prestimulus oscillatory states, we inspected the prestimulus interval (time windows centred between -300 and 0 ms relative to stimulus onset) for potential differences

between SR_{WAIVED} and $SR_{SELECTED}$ but found no evidence for confidence-discriminating activity in any of the frequencies of interest, even at lenient thresholds ($p_{THRESHOLD}$=.05, two-sided test; $p_{CLUSTER}$=.05). To formally assess this, we also compared average condition differences (i.e., power estimates for SR_{WAIVED} minus $SR_{SELECTED}$) between the prestimulus and decision periods (i.e., -300 to 0 ms vs. 50 to 1000 ms, relative to stimulus onset), and showed these were significantly larger during the decision stage of the trial for both frequency ranges of interest ($t(18)=4.51$, $p<.001$, and $t(18)=2.68$, $p=.015$, respectively). Together, these results suggest that observed confidence effects were unlikely to reflect biases carried on from the prestimulus period.

Relationship with non-oscillatory signatures of confidence. In Chapter 2, we performed a temporal characterisation of non-oscillatory (i.e., time-domain) neural signatures of confidence, using a single-trial multivariate classification analysis of the EEG. In short, we identified a transient neural component (referred to as $\boldsymbol{y}$, see Chapter 2 Materials and methods section) which discriminated between SR_{WAIVED} and $SR_{SELECTED}$ trials, beginning early in the trial and peaking approximately 600 ms after stimulus onset (see Fig. 2.2B). Importantly, this confidence-discriminating activity appeared to develop simultaneously with the decision process, and was reflected in the same process of evidence accumulation that characterised it. Moreover, its corresponding scalp topography showed contributions from centroparietal electrodes - this was distinct from the topography associated with confidence-related oscillatory activity, which appeared lateralised over sensorimotor regions, thus likely indicating separate underlying neural generators for the two signals.

Although oscillatory power estimates (particularly in lower frequency bands, e.g. <30 Hz) are by their very nature less temporally precise than their time-domain counterparts, it is interesting to note that the observed confidence-discriminating oscillatory activity in the present study appeared to overlap in time with the confidence-related neural component $\boldsymbol{y}$ identified previously in the time-domain. Specifically, group-level confidence effects within Cluster 1 (alpha/beta) and Cluster 2 (beta) oscillatory activity (Fig. 3.2B) showed peaks at 650 ms and 600 ms post-stimulus, respectively (as reflected by the t-values

resulting from paired comparisons between SR_{WAIVED} and $SR_{SELECTED}$ conditions across time; results not depicted).

To further characterise the relationship between these two confidence-discriminating signals, we investigated their correlation at the single-trial level. To this end, we first extracted single-trial power estimates across all trial groups and performed a decibel transformation on these: $dB = 10\,log_{10}(Power)$. Resulting values were then averaged across the spatio-temporo-spectral dimensions that characterised the two clusters (separately for each cluster). We used linear regression to assess how these values correlated with the single-trial estimates of the confidence-discriminating component $\boldsymbol{y}$, extracted from subject-specific time windows of peak confidence discrimination (see Chapter 2, Materials and Methods section). This analysis was performed separately for each subject. We assessed significance using the non-parametric Wilcoxon signed rank test for non-normally distributed data. As we expected a negative relationship between the two confidence-related signals, we tested whether regression coefficients (betas) resulting across subjects came from a distribution with a median smaller than 0. Indeed, we found a significant effect at the group level (Cluster 1: Z = -3.54, p <. 001; Cluster 2: Z = -3.3, p <. 001). For visualisation, this relationship is displayed in Fig. 3.4.

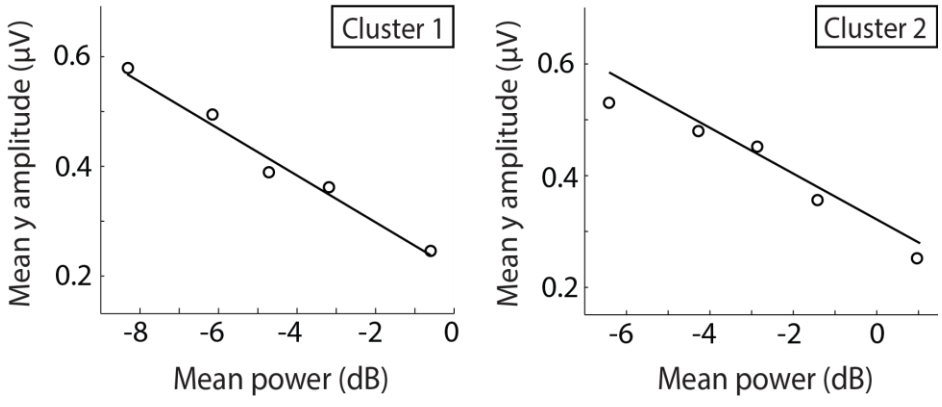

Figure 3.4. Relationship with time-domain confidence signals. Power in the confidence-discriminating oscillatory signals (x-axis) was negatively correlated with the amplitude of the confidence-related component $\boldsymbol{y}$ (y-axis), on a trial-by-trial basis. To visualise this relationship, trials were grouped into five bins based on the magnitude of the power estimates. Power estimates for **A** and **B** are computed using the two data subsets informed by the cluster-based permutation analysis (i.e., Cluster 1 and 2, respectively;

see ‘Frequency analysis’ subsection for details), averaged across the spatial, temporal, and spectral dimensions. Importantly, fitted lines represent fits of the linear regression model to individual trials.

Discussion

Here, we showed that confidence in perceptual judgments was reflected in oscillatory activity within the alpha and beta frequency bands (approx. 10-22 Hz). Specifically, visual stimulation was followed by suppression in the oscillatory power, which was on average more pronounced for high- than low-confidence trials. This effect could be observed during the decision stage of the trial and at least 2 seconds before subjects made a response, and was strongest over the centro-posterior electrodes found contralateral to the motor effector used to express choice (i.e., right hand). We showed that these effects could not be purely explained by the strength of available sensory evidence or by potential spontaneous fluctuations in prestimulus states.

Overall, the spatial and spectral characteristics of these effects appear consistent with a motor-related “desynchronisation”, which refers to the reduction in oscillatory power in the alpha and beta frequency ranges (typically ~8-30 Hz) before and/or during movement, and which is most prominent over sensorimotor regions contralateral to the motor effector (Pfurtscheller and Lopes da Silva 1999). While it is typically associated with execution (Pfurtscheller and Aranibar 1979), planning/preparation (Tzagarakis et al. 2010, Tzagarakis et al. 2015, Pfurtscheller and Berghold 1989, Pfurtscheller and Lopes da Silva 1999) or representation (McFarland et al. 2000) of movement, a few recent studies have also involved alpha and beta suppression in carrying information about action-informative perceptual decisions (Donner et al. 2009, O'Connell et al. 2012, de Lange et al. 2013).

The relationship between confidence and low-frequency neural oscillations observed here is in line with a recent report by Kubanek et al. (2015). In their study, subjects performed an auditory discrimination task, and reported their decision using button presses. Consistent with our results, confident choices showed stronger power suppression in the alpha band following stimulus

presentation, which authors interpret as reflecting confidence in the impending action.

Interestingly, in our experiment, we showed that differences between high- and low-confidence trial groups were detectable in the oscillatory activity as early as 300ms, and peaked approximately 600-650 ms after stimulus onset. The temporal profile of this effect appeared similar to a confidence-related neural component we identified in the time-domain (Gherman and Philiastides, 2015; see Chapter 2, Fig. 2.2B). This latter signal resembled a process of decision-related evidence accumulation and carried information about subjects' confidence. Importantly, its scalp topography differed considerably from the spatial distribution of the confidence-related oscillatory activity, in that it mainly showed contributions from centroparietal sites (as opposed to lateralised central/centroposterior distribution observed in the oscillatory activity). Thus, while correlated, the two signals likely rely on separate neural generators, in line with the observation that motor-preparatory activity is distinct from action-independent decision processes (Kelly and O'Connell 2013, Wyart et al. 2012, Filimon et al. 2013).

Nevertheless, the idea that underlying neural processes for the two different signals may occur in parallel or in temporal proximity is consistent with a growing body of literature suggesting that as the decision forms, decision-related information "leaks" into motor centres that support relevant impending action, thus facilitating efficient response (Song and Nakayama 2009). For instance, electrophysiological work indicates that, when the mapping between a stimulus and motor effector is known, contralateral oscillatory activity in the alpha/beta bands is modulated by the strength of the sensory evidence, can predict choice-related behaviour, and is characterised by a gradual buildup pattern that begins during stimulus viewing, suggesting motor-preparatory signals reflect the evolving decision process (O'Connell et al. 2012, Donner et al. 2009, de Lange et al. 2013, Haegens et al. 2011). Similar findings from animal electrophysiology support this hypothesis, showing that regions of the brain that support choice-relevant action (e.g., lateral intraparietal cortex in the non-human primate brain) encode a gradual build-up of decision-related evidence (de Lafuente et al.

2015, Huk and Shadlen 2005), suggesting that motor-preparatory systems reflect the decision process as it forms, possibly via a continuous flow of information from decision making areas (Gold and Shadlen 2000, Gold and Shadlen 2003, Selen et al. 2012). Notably, decision-related evidence in these regions has also been shown to carry information about eventual confidence in that decision (Kiani and Shadlen 2009). Thus, within this framework, it is possible that the confidence-discriminating oscillatory activity we observe over sensorimotor sites in the present study may reflect the evolving decision process, as well as the confidence-related information it holds.

We did not observe confidence-related activity in the theta frequency band. Midfrontal theta activity is thought to support performance/error monitoring (Murphy et al. 2015, Cohen 2016), a process argued to share a common mechanism with confidence-related processes (Boldt and Yeung 2015, Yeung and Summerfield 2012). However, error monitoring signals are typically observed following overt behaviour during speeded-response tasks (i.e., in response to detected errors), whereas the current task required subjects to wait through a delay prior to making a response, and thus was unlikely to have engaged such a mechanism during the decision phase of the trial.

Similarly, no confidence-related activity was identified in the gamma band, which was previously shown to encode the decision-related evidence accumulation (Polania et al. 2014, Donner et al. 2009) and to discriminate between high- and low-confidence choices (Peters et al. 2017). However, these studies have used MEG and electrocorticography, respectively, which may be better suited for recording gamma activity due to their superior signal-to-noise ratio and spatial sensitivity (Crone et al. 2006, Cheyne 2013).

A potential limitation of the present study was the absence of any measurements of overt motor behaviour during the perceptual decision task. This would be necessary in order to ensure the observed effects could not merely be explained by movement of the motor effector during stimulus presentation and/or decision. We note however that similar effects have been observed in the

absence of motor behaviour, as recorded with electromyography (Kubanek et al. 2015).

To conclude, we showed that, during perceptual decision making, putative motor-systems corresponding to the motor effector appear to store signals that dissociate between subjects' eventual confidence, and do so considerably in advance of overt motor responses. Overall, the temporal profile of this activity appears consistent with a potentially continuous input of decision- and confidence-related information to these regions. Additional research will be needed to understand the mechanisms by which confidence-discriminating signals originate here, specifically whether their link with confidence may be an epiphenomenon of their correlation with the decision variable, or whether the confidence-related information may serve an adaptive function, for example by influencing action itself or potentially even serving as input to further metacognitive evaluation and communication processes.

Chapter 4. Human VMPFC encodes early signatures of confidence in perceptual decisions

Summary

Decision confidence refers to an individual's internal estimate of judgment accuracy, and thus plays a critical role in adaptive behaviour. Correspondingly, recent years have seen significant progress towards understanding its neural basis in relation to post-decisional metacognitive evaluation. Despite this progress however, the early, decisional, stages of confidence processing remain underexplored. Here, we used a simultaneous EEG/fMRI approach to provide a spatiotemporal account of confidence during perceptual decision making. Participants performed a random-dot direction discrimination task and rated their confidence on each trial. Using a multivariate single-trial classifier on the EEG data, we identified a stimulus- and accuracy-independent neural component which discriminated between High vs. Low confidence trials, and which appeared prior to participants' behavioural response. Crucially, we used the trial-to-trial variability of this EEG-derived confidence signal to detect associated fMRI responses in the ventromedial prefrontal cortex (VMPFC), a region not previously linked with confidence for perceptual decisions. Notably, this activation was additional to what could be explained by subjects' confidence ratings alone, and by potential confounding variables (perceptual accuracy, response time, and attention). Our results raise the possibility that the VMPFC supports an early readout of perceptual decision confidence, and are in line with recent work proposing a domain-general role for this region in encoding confidence.

Introduction

Our everyday lives involve frequent situations where we must make judgments based on noisy or incomplete sensory information – for example deciding whether crossing the street on a foggy morning, in poor visibility, is safe. Being able to rely on an internal estimate of whether our perceptual judgments are accurate is fundamental to adaptive behaviour and accordingly, recent years have seen a growing interest in understanding the neural basis of confidence judgments.

Within the perceptual decision making field, one line of research has focused specifically on identifying neural correlates of confidence during metacognitive evaluation (i.e., while subjects actively judge their performance following a choice), and demonstrated the functional involvement of the anterior prefrontal cortex (Fleming et al. 2012, Hilgenstock et al. 2014). Concurrently, psychophysiological work in humans and non-human primates using time-resolved measurements have shown that confidence encoding can also be observed at earlier stages, and as early as the decision process itself (Kiani and Shadlen 2009, Gherman and Philiastides 2015, Zizlsperger et al. 2014).

Correspondingly, recent fMRI studies have reported confidence-related signals nearer the time of decision (e.g., during perceptual stimulation) in regions such as the striatum (Hebart et al. 2016), dorsomedial prefrontal cortex (Heereman et al. 2015), cingulate and insular cortices (Paul et al. 2015), and other areas of the prefrontal, parietal, and occipital cortices (Heereman et al. 2015, Paul et al. 2015). Interestingly, confidence-related processing has also been reported in the ventromedial prefrontal cortex (VMPFC) during value-based and a range of ratings tasks (De Martino et al., 2013; Lebreton et al., 2015), however the extent to which this region is additionally involved in perceptual judgments relying on temporal integration of sensory evidence remains unclear.

Importantly, research investigating the neural correlates of decision confidence has thus far relied – nearly exclusively – on correlations with behavioural measures, the most common of these being the subjective ratings given by participants after the decision (see Grimaldi et al., 2015, for a review).

However, theoretical and empirical work suggests that post-decisional metacognitive judgments may be affected by processes occurring after termination of the initial decision (Fleming et al. 2015, Moran et al. 2015, Pleskac and Busemeyer 2010, Yu et al. 2015, Murphy et al. 2015, Fleming and Daw 2017, van den Berg et al. 2016a, Navajas et al. 2016), such as integration of existing information, processing of novel information arriving post-decisionally, or decay (Moran et al. 2015), and may consequently be only partly reflective of early confidence-related states.

Here we aim to derive a more faithful representation of these early confidence signals using EEG, and exploit the trial-by-trial variability in these signals to build parametric EEG-informed fMRI predictors, thus aiming to provide a more complete spatiotemporal account of decision confidence. We hypothesise that using an electrophysiologically-derived (i.e. endogenous) representation of confidence to detect associated fMRI responses would provide not only a more temporally precise, but also a more accurate spatial representation of confidence around the time of decision.

To test this hypothesis, we collected simultaneous EEG/fMRI data while participants performed a random-dot direction discrimination task and rated their confidence on each trial. Using a multivariate single-trial classifier to discriminate between High vs. Low confidence trials in the EEG data, we extracted an early, stimulus- and accuracy-independent discriminant component appearing prior to participants' behavioural response. We then regressed the resultant single-trial component amplitudes against the fMRI signal and identified a positive correlation with this early confidence signal in a region of the VMPFC that has not been previously linked to perceptual decisions. Crucially, activation of this region was unique to our EEG-informed fMRI predictor (i.e., additional to those detected with a conventional fMRI regressor, which relied solely on participants' post-decisional confidence reports).

Materials and Methods

Participants. Thirty subjects participated in the simultaneous EEG/fMRI experiment. Four were subsequently removed from the analysis due to near chance (n=3) and ceiling (n=1) performance, respectively, on the perceptual discrimination task. Additionally, one subject was excluded whose confidence reports covered only a limited fraction of the provided rating scale, thus yielding an insufficient number of trials to be used in the EEG discrimination analysis (see below). Finally, one subject had to be removed due to poor (chance) performance of the EEG decoder (see below). All results presented here are based on the remaining 24 subjects (age range 20-32 years). All were right-handed, had normal or corrected to normal vision, and reported no history of neurological problems. The study was approved by the College of Science and Engineering Ethics Committee at the University of Glasgow (CSE01355) and informed consent was obtained from all participants.

Stimuli and task. All stimuli were created and presented using the PsychoPy software (Peirce 2007). They were displayed via an LCD projector (frame rate=60Hz) on a screen placed at the rear opening of the bore of the MRI scanner, and viewed through a mirror mounted on the head coil (distance to screen = 95cm). Stimuli consisted of random dot kinematograms (Newsome and Pare 1988), whereby a proportion of the dots moved coherently to one direction (left vs. right), while the remainder of the dots moved at random. Specifically, each stimulus consisted of a dynamic field of white dots (number of dots=150; dot diameter=0.1 degrees of visual angle, dva; dot life time=4 frames; dot speed=6 dva/s), displayed centrally on a grey background through a circular aperture (diameter=6 dva). Task difficulty was controlled by manipulating the proportion of dots moving coherently in the same direction (i.e., motion coherence).

We aimed to maintain overall performance on the main perceptual decision task consistent across subjects (i.e., near perceptual threshold, at approximately 75% correct). For this reason, task difficulty was calibrated individually for each

subject on the basis of a separate training session, prior to the day of the main experiment.

Training. To first familiarise subjects with the random dot stimuli and facilitate learning on the motion discrimination task, subjects first performed a short simplified version of the main task (lasting approx. 10 minutes), where feedback was provided on each trial. The task, which required making speeded direction discriminations of random dot stimuli (see below), began at a low-difficulty level (motion coherence = 40%) and gradually increased in difficulty in accordance with subjects' online behavioural performance (a 3-down-1-up staircase procedure, where three consecutive correct responses resulted in a 5% decrease in motion coherence, whereas one incorrect response yielded a 5% increase). This was followed by a second, similar task, which served to determine subject-specific psychophysical thresholds. Seven motion coherence levels (5%, 8%, 12%, 18%, 28%, 44%, 70%) were equally and randomly distributed across 350 trials. The proportion of correct responses was separately computed for each motion coherence level, and a logarithmic function was fitted through the resulting values in order to estimate an optimal motion coherence yielding a mean performance of approximately 75% correct. Subjects who showed near-chance performance across all coherence levels or showed no improvement in performance with increasing motion coherence were not tested further and did not participate in the main experiment. No feedback was given for this or any of the subsequent tasks.

Main task. On the day of the main experiment, subjects practised the main task once outside the scanner, and again inside the scanner prior to the start of the scan (a short 80 trial block each time). The main task required subjects to judge the motion direction of random dot kinematograms (left vs. right) and rate how confident they were in their choice, on a trial-by-trial basis (Fig. 2.1A).

Each trial began with a random dot stimulus lasting for a maximum of 1.2 s, or until the subject made a behavioural response. Subjects were instructed to respond as quickly as possible, and had a time limit of 1.35 s to do so. The message "Oops! Too slow" was displayed if this time limit was exceeded or no direction response was made. Once the dot stimulus disappeared, the screen

remained blank until the 1.2 s stimulation period elapsed and through an additional random delay (1.5-4 s). Next, subjects were presented with a rating scale for 3 s, during which they reported their confidence in the previous direction decision. The confidence scale was represented intuitively by means of a white horizontal bar of linearly varying thickness, with the thick end representing high confidence. Its orientation on the horizontal axis (thin-to-thick vs. thick-to-thin) informed subjects of the response mapping, and this was equally and randomly distributed across trials to control for motor preparation effects. To make a confidence response, subjects moved an indicator (a small white triangle) along a 9-point marked line. The indicator changed colour from white to yellow when a confidence response was selected and this remained on the screen until the 3 s elapsed). A final delay (blank screen, jittered between 1.5-4 s) ended the trial. Failing to provide either a direction or a confidence response within the respective allocated time limits on a given trial rendered it invalid, and this was subsequently removed from further analyses. This resulted in a total fraction of .04 (.02 and .02, respectively) of trials being discarded.

Subjects performed 2 experimental blocks of 160 trials each, corresponding to two separate fMRI runs. Each block contained two short (30 s) rest breaks, during which the MR scanner continued to run. Subjects were instructed to remain still throughout the entire duration of the experiment, including during rest breaks and in between scans. Motion coherence was held constant across trials, at the subject-specific level estimated during training. The direction of the dots was equally and randomly distributed across trials. To control for confounding effects of low-level trial-to-trial variability in stimulus properties on decision confidence, an identical set of stimuli was used in the two experimental blocks. Specifically, for each subject, the random seed, which controlled dot stimulus motion parameters in the stimulus presentation software was set to a fixed value. This manipulation allowed for subsequent control comparisons between pairs of identical stimuli.

Subjects were encouraged to explore the entire scale when making their responses and to abstain from making a confidence response on a given trial if they became aware of having made a motor mapping error. This was in an effort

to minimise the influence of premature responses, which are likely caused by a release the motor system from global inhibition under time pressure (Forstmann et al. 2008), and therefore unreflective of the decision process or confidence per se. Additionally, subjects were instructed to make their responses as quickly and accurately as possible, and provide a response on every trial. All behavioural responses were executed using the right hand, on an MR-compatible button box.

EEG data acquisition. EEG data was collected simultaneously with the fMRI data during performance of the main task, using an MR-compatible EEG amplifier system (Brain Products, Germany). Continuous EEG data was recorded using the Brain Vision Recorder software (Brain Products, Germany) at a sampling rate of 5000 Hz. We used 64 Ag/AgCl scalp electrodes positioned according to the 10-20 system, and one nasion electrode. Reference and ground electrodes were embedded in the EEG cap and were located along the midline, between electrodes Fpz and Fz, and between electrodes Pz and Oz, respectively. Each electrode had in-line 10 kOhm surface-mount resistors to ensure subject safety. Input impedance was adjusted to <25 kOhm for all electrodes. Acquisition of the EEG data was synchronized with the MR data acquisition (Syncbox, Brain Products, Germany), and MR-scanner triggers were collected separately to enable offline removal of MR gradient artifacts from the EEG signal. Scanner trigger pulses were lengthened to 50μs using a built-in pulse stretcher, to facilitate accurate capture by the recording software. Experimental event markers (including participants' responses) were synchronized, and recorded simultaneously, with the EEG data.

EEG data processing. Preprocessing of the EEG signals was performed using Matlab (Mathworks, Natick, MA). EEG signals recorded inside an MR scanner are contaminated with gradient artifacts and ballistocardiogram (BCG) artifacts due to magnetic induction on the EEG leads. To correct for gradient-related artifacts, we constructed average artifact templates from sets of 80 consecutive functional volumes centred on each volume of interest, and subtracted these from the EEG signal. This process was repeated for each functional volume in our dataset. Additionally, a 12 ms median filter was applied in order to remove any residual spike artifacts. Further, we corrected for standard EEG artifacts and

applied a 0.5–40 Hz band-pass filter in order to remove slow DC drifts and high frequency noise. All data were downsampled to 1000 Hz.

To remove eye movement artifacts, subjects performed an eye movement calibration task prior to the main experiment (with the MRI scanner turned off, to avoid gradient artifacts), during which they were instructed to blink repeatedly several times while a central fixation cross was displayed in the centre of the computer screen, and to make lateral and vertical saccades according to the position of the fixation cross. We recorded the timing of these visual cues and used principal component analysis to identify linear components associated with blinks and saccades, which were subsequently removed from the EEG data (Parra et al. 2005).

Next, we corrected for cardiac-related (i.e., ballistocardiogram, BCG) artifacts. As these share frequency content with the EEG, they are more challenging to remove. To minimise loss of signal power in the underlying EEG signal, we adopted a conservative approach by only removing a small number of subject-specific BCG components, using principal component analysis. We relied on the single-trial classifiers to identify discriminating components that are likely to be orthogonal to the BCG. BCG principal components were extracted from the data after the data were first low-pass filtered at 4 Hz to extract the signal within the frequency range where BCG artifacts are observed. Subject-specific principal components were then determined (average number of components across subjects: 1.8). The sensor weightings corresponding to those components were projected onto the broadband data and subtracted out. Finally, data were baseline corrected by removing the average signal during the 100 ms prestimulus interval.

Single-trial EEG analysis. To identify confidence-related signals in the EEG data with increased statistical power, we first separated trials into three confidence groups (Low, Medium, High), on the basis of the original 9-point confidence rating scale. Specifically, we isolated High- and Low-confidence trials by pooling across each subject's three highest and three lowest ratings, respectively. To ensure robustness of our single trial EEG analysis, we imposed a minimum limit of 50 trials per confidence trial group. For those data sets where subjects had an

insufficient number of trials in the extreme ends of the confidence scale, neighbouring confidence bins were included to meet this limit.

We used a single-trial multivariate discriminant analysis, combined with a sliding window approach (Parra et al. 2005, Sajda et al. 2009) to discriminate between High and Low confidence trials in the stimulus-locked EEG data. This method aims to estimate, for predefined time windows of interest, an optimal combination of EEG sensor linear weights (i.e., a spatial filter) which, applied to the multichannel EEG data, yields a one-dimensional projection (i.e., a "discriminant component") that maximally discriminates between the two conditions of interest. Importantly, unlike univariate trial-average approaches for event-related potential analysis, this method spatially integrates information across the multidimensional sensor space, thus increasing signal-to-noise ratio whilst simultaneously preserving the trial-by-trial variability in the signal, which may contain task-relevant information. In our data, we identified confidence-related discriminating components, $\boldsymbol{y}(t)$, by applying a spatial weighting vector $\boldsymbol{w}$ to our multidimensional EEG data $\boldsymbol{x}(t)$, as follows:

$$\boldsymbol{y}(t) = \boldsymbol{w}^T \boldsymbol{x}(t) = \sum_{i=1}^{D} w_i x_i(t) \quad (1)$$

where D represents the number of channels, indexed by i, and T indicates the transpose of the matrix. To estimate the optimal discriminating spatial weighting vector $\boldsymbol{w}$, we used logistic regression and a reweighted least squares algorithm (Jordan and Jacobs 1994). We applied this method to identify $\boldsymbol{w}$ for short (60 ms) overlapping time windows centred at 10 ms-interval time points, between -100 and 1000 ms relative to the onset of the random dot stimulus (i.e., the perceptual decision phase of the trial). This procedure was repeated for each subject and time window. Applied to an individual trial, spatial filters ($\boldsymbol{w}$) obtained this way produce a measurement of the discriminant component amplitude for that trial. In separating the High and Low trial groups, the discriminator was designed to map the component amplitudes for one condition to positive values and those of the other condition to negative values. Here, we mapped the High confidence trials to positive values and the Low confidence trials to negative values, however note that this mapping is arbitrary.

To quantify the performance of the discriminator for each time window, we computed the area under a receiver operating characteristic (ROC) curve (i.e., the Az value), using a leave-one-out trial procedure (Duda et al. 2001). We determined significance thresholds for the discriminator performance using a bootstrap analysis whereby trial labels were randomised and submitted to a leave-one-out test. This randomisation procedure was repeated 500 times, producing a probability distribution for *Az*, which we used as reference to estimate the *Az* value leading to a significance level of $p<0.01$.

Given the linearity of our model we also computed scalp projections of the discriminating components resulting from Eq. 1 by estimating a forward model for each component:

$$\mathbf{a} = \frac{\boldsymbol{X}\,\boldsymbol{y}}{\boldsymbol{y}^T\boldsymbol{y}} \qquad (2)$$

where the EEG data ($\boldsymbol{X}$) and discriminating components ($\boldsymbol{y}$) are now in a matrix and vector notation, respectively, for convenience (i.e., both $\boldsymbol{X}$ and $\boldsymbol{y}$ now contain a time dimension). Equation 2 describes the electrical coupling of the discriminating component $\boldsymbol{y}$ that explains most of the activity in $\boldsymbol{X}$. Strong coupling indicates low attenuation of the component $\boldsymbol{y}$ and can be visualised as the intensity of vector **a**.

Single-trial power analysis. We calculated prestimulus alpha power (8-12Hz) in the 400 ms epoch beginning at -500 ms relative to the onset of the random dot stimulus. To do this, we used the multitaper method (Mitra and Pesaran 1999) as implemented in the FieldTrip toolbox for Matlab (http://www.ru.nl/neuroimaging/fieldtrip). Specifically, for each epoch data were tapered using discrete prolate spheroidal sequences (2 tapers for each epoch; frequency smoothing of ±4Hz) and Fourier transformed. Resulting frequency representations were averaged across tapers and frequencies. Single-trial power estimates were then extracted from the occipitoparietal sensor with

the highest overall alpha power and baseline normalised through conversion to decibel units (dB).

MRI data acquisition. Imaging was performed at the Centre for Cognitive Neuroimaging, Glasgow, using a 3-Tesla Siemens TIM Trio MRI scanner (Siemens, Erlangen, Germany) with a 12-channel head coil. Cushions were placed around the head to minimize head motion. We recorded two experimental runs of 794 whole-brain volumes each, corresponding to the two blocks of trials in the main experimental task. Functional volumes were acquired using a T2*-weighted gradient echo, echo-planar imaging sequence (32 interleaved slices, gap: 0.3 mm, voxel size: 3 × 3 × 3 mm, matrix size: 70 × 70, FOV: 210 mm, TE: 30 ms, TR: 2000 ms, flip angle: 80°). Additionally, a high-resolution anatomical volume was acquired at the end of the experimental session using a T1-weighted sequence (192 slices, gap: 0.5 mm, voxel size: 1 × 1 × 1 mm, matrix size: 256 × 256, FOV: 256 mm, TE: 2300 ms, TR: 2.96 ms, flip angle: 9°), which served as anatomical reference for the functional scans.

fMRI preprocessing. The first 10 volumes prior to task onset were discarded from each fMRI run to ensure a steady-state MR signal. Additionally, 13 volumes were discarded from the post-task period at the end of each block. The remaining 771 volumes were used for statistical analyses. Pre-processing of the MRI data was performed using the FEAT tool of the FSL software (http://www.fmrib.ox.ac.uk/fsl) and included slice-timing correction, high-pass filtering (>100 s), and spatial smoothing (with a Gaussian kernel of 8 mm full width at half maximum), and head motion correction (using the MCFLIRT tool). The motion correction preprocessing step generated motion parameters which were subsequently included as regressors of no interest in the general linear model (GLM) analysis (see fMRI analysis below). Brain extraction of the structural and functional images was performed using the Brain Extraction tool (BET). Registration of EPI images to standard space (Montreal Neurological Institute, MNI) was performed using the Non-linear Image Registration Tool with a 10-mm warp resolution. The registration procedure involved transforming the EPI images into an individual's high-resolution space (with a linear, boundary-based registration algorithm, (Greve and Fischl 2009)) prior to transforming to standard

space. Registration outcome was visually checked for each subject to ensure correct alignment.

fMRI analysis. Whole-brain statistical analyses of functional data were conducted using a general linear model (GLM) approach, as implemented in FSL (FEAT tool):

$$Y = \beta X + \varepsilon = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \quad (3)$$

where Y represents the BOLD response time series for a given voxel, structured as a T×1 (T time samples) column vector, and X represents the T×N (N regressors) design matrix, with each column representing one of the psychological regressors (see GLM analysis below for details), convolved with a canonical hemodynamic response function (double-gamma function). β represents the parameter estimates (i.e., regressor betas) resulting from the GLM analysis in the form of a N × 1 column vector. Lastly, ε is a T × 1 column vector of residual error terms. A first-level analysis was performed to analyse each subject's individual runs. These were then combined at the subject-level using a second-level analysis (fixed effects). Finally, a third-level mixed-effects model (FLAME 1) was used to combine data across all subjects.

Simultaneous EEG/fMRI analysis. With the combined EEG/fMRI approach, we sought to identify confidence-related activation in the fMRI surpassing what could be explained by the relevant behavioural predictors alone. In particular, we looked for brain regions where BOLD responses correlated with the confidence-discriminating component derived from the EEG analysis. Our primary motivation behind this approach was the hypothesis that endogenous trial-by-trial variability in the confidence discriminating EEG component (near the time of perceptual decision, and prior to behavioural response) would be more reflective of early internal representations of confidence at the single-trial level, compared to the metacognitive reports which are provided post-decisionally and therefore likely to be subjected to additional processes. We predicted that the simultaneous EEG/fMRI approach would enable identification of latent brain states that might remain unobserved with a conventional analysis

approach. To this end, we extracted trial-by-trial amplitudes of $\boldsymbol{y}(t)$ (resulting from Eq. 1) at the time window of maximum confidence discrimination, and used these to build a BOLD predictor, which we henceforth refer to as the Y_{CONF} regressor. Importantly, to avoid possible confounding effects of motor preparation/response, the time of this component was determined on a subject-specific basis, by only considering the period prior to the behavioural choice (mean peak discrimination time = 708 ms from stimulus onset, SD=162 ms). Thus, on average this was selected 287ms (SD=171 ms) prior to each subject's mean response time.

Note that the trial-by-trial variability in our EEG component amplitudes is driven mostly by cortical regions found in close proximity to the recording sensors and to a lesser extent by distant (e.g., subcortical) structures. Nonetheless, an advantage of our EEG-informed fMRI predictors is that they can also reveal relevant fMRI activations within deeper structures, provided that their BOLD activity covaries with that of the cortical sources of our EEG signal.

GLM analysis. We designed our GLM model to account for variance in the BOLD signal at two key stages of the trial, namely the perceptual decision period (beginning at the onset of the random dot visual stimulus) and the metacognitive evaluation/rating (beginning at the onset of the rating scale display), respectively. A total of 10 regressors were included in the model. Our primary predictor of interest was the EEG-derived endogenous measure of confidence (Y_{CONF} regressor). We modelled this as a stick function (duration = 0.1 s) locked to the stimulus onset, with event amplitudes parametrically modulated by the trial-to-trial variability in the confidence discriminating component $\boldsymbol{y}(t)$. To ensure variance explained by this regressor was unique (i.e., not explained by subjects' behavioural reports), we included a second regressor whose event amplitudes were parametrically modulated by confidence ratings, and which was otherwise identical to the Y_{CONF} regressor (i.e., $Ratings_{DEC}$ regressor, duration = 0.1 s, locked to stimulus onset). Importantly, Y_{CONF} amplitudes were only moderately correlated with behavioural confidence ratings (mean R=.39, SD=.07), thus allowing us to exploit additional explanatory power inherent to this regressor. Other regressors of no interest for the perceptual decision stage

included: one regressor parametrically modulated by prestimulus alpha power in the EEG signal (to control for potential attentional baseline effects), one categorical regressor (1/0) accounting for variability in response accuracy, and one unmodulated regressor (all event amplitudes set to 1) modelling stimulus-related visual responses of no interest across both valid and non-valid (missed) trials (all event durations = 0.1 s, locked to stimulus onset). To control for motor preparation/response, we also included a parametric regressor modulated by subjects' reaction time on the direction discrimination task (duration = 0.1 s, locked to the time of behavioural response).

Additionally, locked to the onset of the metacognitive rating period, we included one parametric regressor (duration = 0.1 s) with event amplitudes modulated by subjects' confidence ratings, one boxcar regressor with duration equivalent to subjects' active behavioural engagement in confidence rating (to minimise effects relating to motor processes), and one unmodulated regressor (duration = 0.1 s). Lastly, we included one categorical boxcar regressor (1/0) to model non-task activation (i.e., rest breaks within each run). Motion correction parameters obtained from fMRI preprocessing were entered as additional covariates of no interest.

Resampling procedure for fMRI thresholding. To estimate a significance threshold for our fMRI statistical maps whilst correcting for multiple comparisons, we performed a nonparametric permutation analysis that took into account the a priori statistics of the trial-to-trial variability in our primary regressor of interest (Y_{CONF}), in a way that trades off cluster size and maximum voxel Z-score (Debettencourt et al. 2011). For each resampled iteration, we maintained the onset and duration of the regressor identical, whilst shuffling amplitude values across trials, runs and subjects. Thus, the resulting regressors for each subject were different as they were constructed from a random sequence of regressor amplitude events. This procedure was repeated 200 times. For each of the 200 resampled iterations, we performed a full 3-level analysis (run, subject, and group). Our design matrix included the same regressors of non-interest used in all our GLM analysis. This allowed us to construct the null hypothesis H_0, and establish a threshold on cluster size and Z-score based on the

cluster outputs from the permuted parametric regressors. Specifically, we extracted cluster sizes from all activations exceeding a minimal cluster size (5 voxels) and Z-score (2.57 per voxel) for positive correlations with the permuted parametric regressors. Finally, we examined the distribution of cluster sizes (number of voxels) for the permuted data and found that the largest 5% of cluster sizes exceeded 162 voxels. We therefore used these results to derive a corrected threshold for our statistical maps, which we then applied to the clusters observed in the original data (that is, Z=2.57, minimum cluster size of 162 voxels, corrected at p=0.05).

Results

Behaviour. On average, subjects indicated their decision on the direction discrimination task 994 ms (SD = 172 ms) after stimulus onset and, consistent with our subject-specific calibrations of the stimulus difficulty (i.e., targeting psychophysical threshold), they performed correctly on 75% (SD = 5.2%) of the trials. In providing behavioural confidence reports, subjects tended to employ the entire rating scale, showing that subjective confidence varied from trial-to-trial despite perceptual evidence remaining constant throughout the task. As a general measure of validity of subjects' confidence reports, we first examined the relationship with behavioural task performance. Specifically, confidence is largely known to scale positively with decision accuracy and negatively with response time (Vickers and Packer 1982, Baranski and Petrusic 1998) (though this relationship is not perfect, and is subject to individual differences, e.g., (Fleming and Dolan 2012, Fleming et al. 2010, Baranski and Petrusic 1994, Zylberberg et al. 2014)). As expected, we found a positive correlation with accuracy (subject-averaged R = .30; one-sample t-test, $t(23) = 13.9$, $p < .001$) (Fig. 4.1B), and a negative correlation with response time (subject-averaged R = -.27; one-sample t-test, $t(23) = -7.8$, $p < .001$) (Fig. 4.1C). Thus, subjects' confidence ratings were generally reflective of their performance on the perceptual decision task.

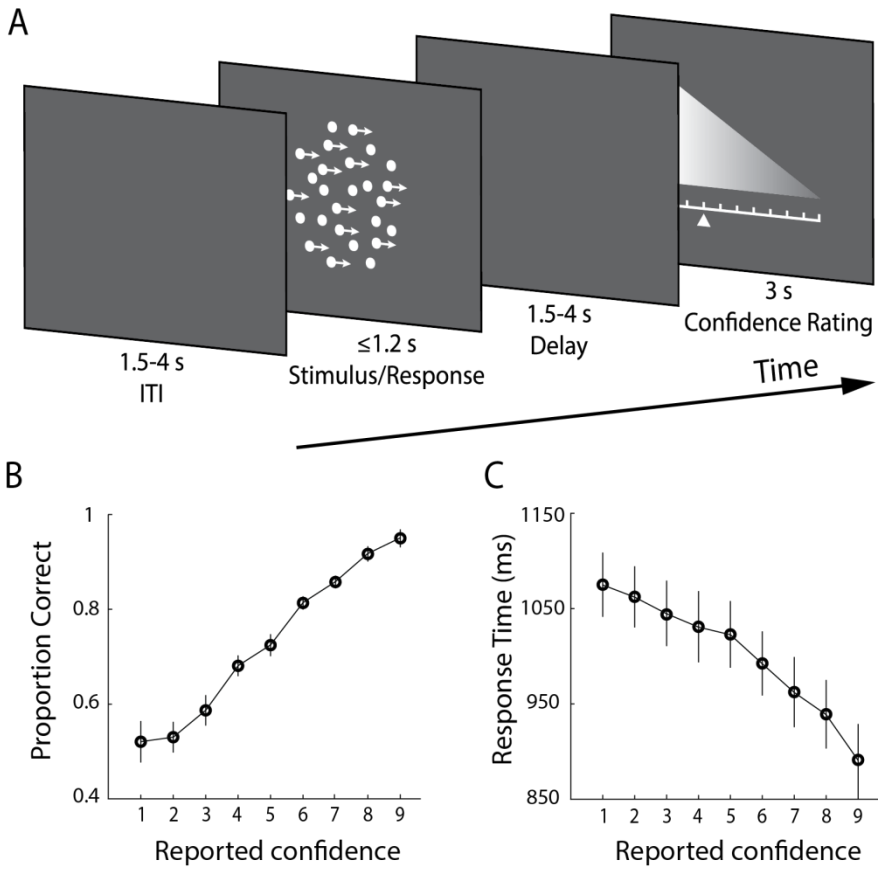

Figure 4.1. Experimental design and behavioural performance. **A.** Schematic representation of the behavioral paradigm. Subjects made speeded left vs. right motion discriminations of random dot kinematograms calibrated to each individual's perceptual threshold. Stimulus difficulty (i.e., motion coherence) and was held constant across trials. Stimuli were presented for up to 1.2 s, or until a behavioural response was made. After each direction decision, subjects rated their confidence on a 9-point scale (3 s). The response mapping for high vs. low confidence ratings alternated randomly across trials to control for motor preparation effects, and was indicated by the horizontal position of the scale, with the tall end representing high confidence. All behavioural responses were made on a button box, using the right hand. **B.** Mean proportion of correct direction choices as a function of reported confidence. **C.** Mean response time as a function of reported confidence. Error bars in B and C represent the standard errors across subjects.

Next, we asked whether subjects' confidence reports could be explained by local fluctuations in attention. To address this possibility, we performed a serial autocorrelation regression analysis on a single subject basis, which predicted confidence ratings on the current trial from ratings given on the immediately preceding five trials. On average, this model accounted for only a minimal

fraction of the variance in confidence ratings (subject-averaged R^2 = .07). Finally, we sought to rule out the possibility that trial-to-trial variability in confidence could be explained by potential subtle differences in low-level physical properties of the stimulus that may go beyond motion coherence (e.g., location and/or timing of individual dots). To this end, we compared subjects' confidence reports on the two experimental blocks which contained an identical set of stimuli, and found no significant correlation between these (R = 0.02, p = 0.44). Taken together, these results support the hypothesis that subjects' reports reflected internal fluctuations in their sense of confidence, which are largely unaccounted for by external factors.

EEG-derived measure of confidence. We conducted a single-trial multivariate discrimination analysis on the EEG data between Low vs. High confidence trials (see Materials and Methods), on the basis of subjective confidence reports. It is important to note that separating trials in this manner only served to increase the precision of the discrimination process, i.e., estimate the electrode contribution patterns that optimally captured confidence. Data from all trials, including those not originally used in the discrimination analysis, were subsequently subjected through these spatial filters, resulting in discriminant component amplitudes that represent graded (individual trial) measures of internal confidence.

We found that discrimination performance (Az) between the two confidence trial groups peaked, on average, 708 ms after stimulus onset (SD = 162ms, Fig. 4.2A). To visualise the spatial extent of this confidence component, we computed a forward model of the discriminating activity (Eq. 2), which can be represented as a scalp map (Fig. 4.2A). Importantly, both the temporal profile and electrode distribution of confidence-related discriminating activity appeared consistent with our previous work (Gherman and Philiastides 2015) where we used stand-alone EEG to identify time-resolved signatures of confidence during a face vs. car task. Together these observations are an indication that the temporal dynamics of decision confidence can be reliably captured using EEG data acquired inside the MR scanner, and that these early confidence-related signals may generalise across tasks.

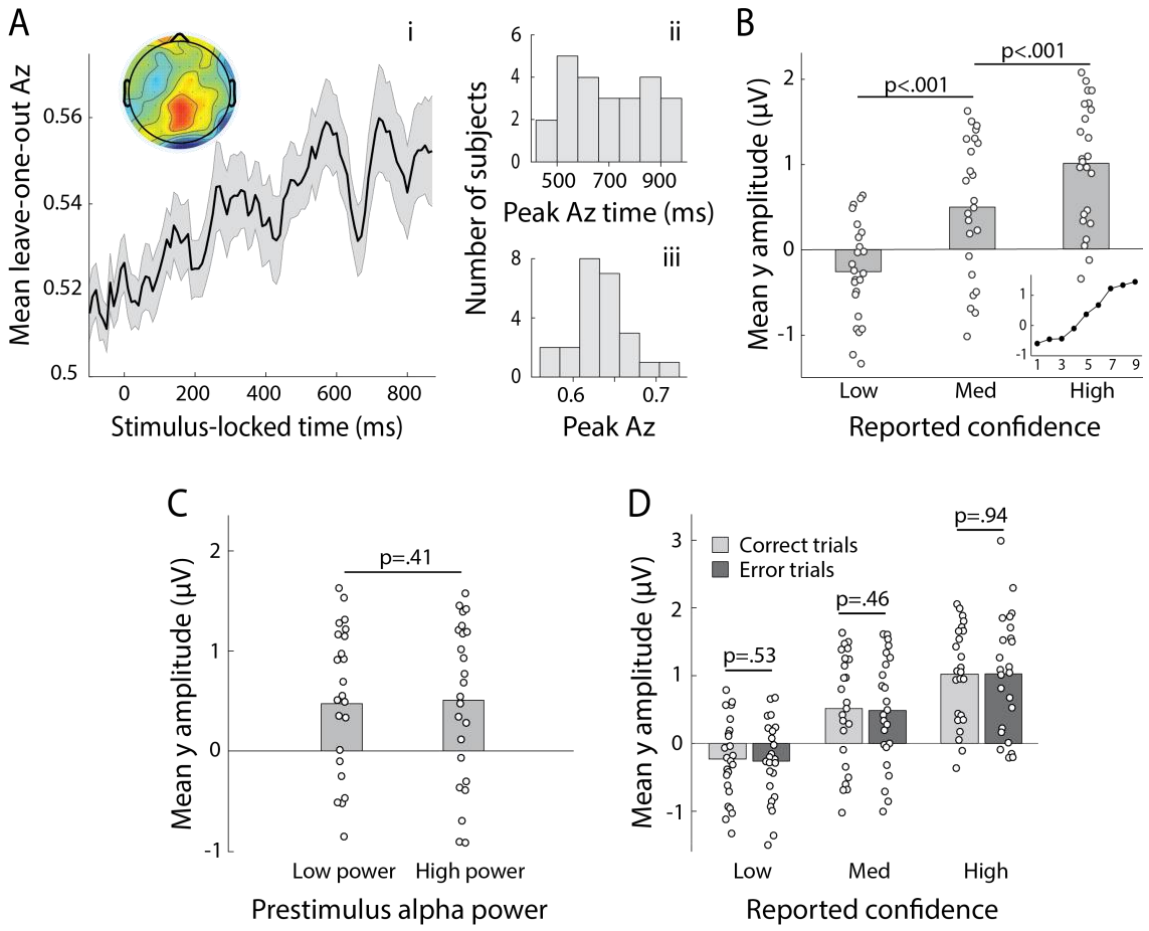

Figure 4.2. Neural representation of confidence in the EEG. **A**. Classifier performance (Az) during High- vs. Low-confidence discrimination for stimulus-locked single-trial data, i. Mean confidence discrimination performance as a function of time (shaded area represents standard errors across subjects). Inset shows average (normalised) topography associated with the discriminating component at subject-specific times of peak confidence discrimination, ii. Distribution of peak confidence discrimination times across subjects. In selecting these, we considered only the discrimination period ending on average at least 100 ms (across-subject mean 271±162 ms) prior to the subjects' mean response times, to minimise potential confounds with activity related to motor execution (due to a sudden increase in corticospinal excitability in this period (Chen et al. 1998), iii. Distribution of Az values at the time of peak confidence discrimination across subjects. **B**. Mean amplitude of the confidence discriminant component as a function of confidence group (Low, Medium, High; grey bars). As expected, component amplitudes for the Medium confidence trials (i.e., trials which were independent from those used to perform the discrimination analysis) are situated between the Low and High confidence trials. The mean component amplitudes for individual confidence ratings (weighted by each subjects' trial count per rating) are also shown (inset). **C**. Mean amplitudes of the confidence discriminant component did not differ significantly between trials showing High vs. Low prestimulus oscillatory power in the alpha band. **D**. Mean amplitude of the confidence discriminant component, separated by confidence group and accuracy on the perceptual task. No significant differences were observed between correct and error trials (light and dark grey bars, respectively). White dots in B, C, and D represent individual subject means.

To provide additional support linking this discriminating component to choice confidence, we considered the Medium-confidence trials. Importantly, these trials can be regarded as "unseen" data, as they are independent from those used to train the classifier. We subjected these trials through the same neural generators (i.e. spatial projections) estimated during discrimination of High vs. Low confidence trials and, as expected from a graded quantity, found that the mean component amplitudes for Medium-confidence trials were situated between, and significantly different from, those in the High- and Low-confidence trial groups (both $p < .001$, Fig. 4.2B).

Further, we aimed to address potential confounding effects in our EEG results. As with the behavioural data, we first addressed the possibility that the observed variability in the confidence discriminating component could be attributed to local fluctuations in attention, by conducting a serial autocorrelation analysis. As before, this model only explained a small fraction of the variance in component amplitudes (subject-averaged $R^2 = .03$). We also assessed the influence of a neural signal known to correlate with attention (Thut et al. 2006) and predict visual discrimination (van Dijk et al. 2008), namely occipitoparietal prestimulus alpha power. To do this, we separated trials into High vs. Low alpha power groups, individually for each subject, and compared the corresponding average discriminant component amplitudes. We found that these did not differ significantly between the two groups (paired t-test, $p=.19$, Fig. 4.2C). Next, we tested whether our results could be explained by subjects' task performance (i.e., accuracy of the direction decision), by comparing discriminant component amplitudes for correct vs. error responses. We performed this comparison separately for each of the three confidence trial groups (Low, Medium, High) and found no significant differences between these (paired t-tests, all $p>.45$, Fig. 4.2D). Finally, we note that variability in the confidence discriminant component was also independent of stimulus difficulty, as this was held constant across all trials. We further supported this by showing that discriminant component amplitudes between the two identical-stimulus experimental blocks were not significantly correlated (mean $R = .02$; one-sample t-test, $p = .39$).

fMRI correlates of reported confidence. Although the fMRI model employed here was primarily aimed at identifying activation correlating with endogenous (electrophysiologically-derived) signatures of confidence at the time of decision, our design matrix also included regressors accounting for variance linked to subjects' behavioural confidence reports, as well as other potentially confounding factors (task performance, response time, attention, and visual stimulation; see Materials and Methods).

Thus, we first inspected the activation patterns associated with confidence ratings during the perceptual decision phase of the trial (Fig. 4.3A). The coordinates of all activations are listed in Table 4.1. We found that the BOLD response increased with reported confidence in the striatum, lateral orbitofrontal cortex (OFC), the ventral anterior cingulate cortex (ACC) – areas thought to play a role in human valuation and reward (Grabenhorst and Rolls 2011, Rushworth et al. 2007, O'Doherty 2004) – as well as the right anterior middle frontal gyrus, amygdala/hippocampus, and visual association areas. Overall, these activations appear consistent with findings from previous studies that have identified spatial correlates of decision confidence (De Martino et al. 2013, Hebart et al. 2016, Heereman et al. 2015, Rolls et al. 2010a). Negative activations (i.e., regions showing increasing BOLD response with decreasing reported confidence) were found in the right supplementary motor area, dorsomedial prefrontal cortex, right inferior frontal gyrus (IFG), anterior insula/frontal operculum, in line with previous reports of decision uncertainty near the time of decision (Hebart et al. 2016, Heereman et al. 2015).

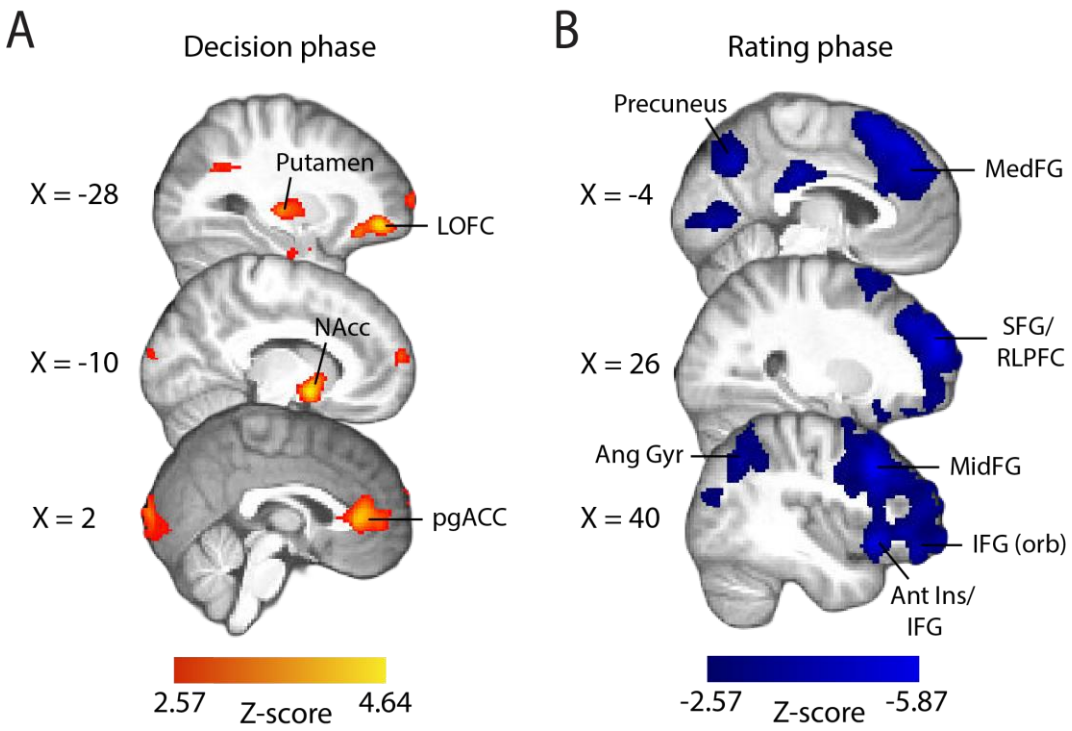

Figure 4.3. Parametric modulation of the BOLD signal by reported confidence. **A.** Clusters showing positive correlation with confidence during the decision phase of the trial. **B.** Clusters showing negative correlation with confidence at the onset of the rating cue (i.e., rating phase). All results are reported at $|Z| \geq 2.57$, and cluster-corrected using a resampling procedure (minimum cluster size 162 voxels; see Materials and Methods). *Ang Gyr*, angular gyrus; *Ant Ins*, anterior insula; *IFG (orb)*, inferior frontal gyrus (orbital region); *LOFC*, lateral orbitofrontal cortex; *MedFG*, medial frontal gyrus; *MidFG*, middle frontal gyrus; *NAcc*, nucleus accumbens; *pgACC*, pregenual anterior cingulate cortex; *RLPFC*, rostrolateral prefrontal cortex; *SFG*, superior frontal gyrus. The complete lists of activations are shown in Tables 1 and 2.

During the metacognitive report stage of the trial (i.e., rating phase, Fig. 4.3B), we found negative correlations with confidence ratings in extended networks (Table 4.2) which included regions of the rostrolateral prefrontal cortex (bilateral, right lateralised), middle frontal gyrus, superior frontal gyrus (extending along the cortical midline and into the medial prefrontal cortex), orbital regions of the IFG, angular gyrus, precuneus, posterior cingulate cortex (PCC), and regions of the occipital and middle temporal cortices. These activations are largely in line with research on the spatial correlates of choice uncertainty (Grinband et al. 2006, Fleming et al. 2012) and metacognitive evaluation (Fleming et al. 2012, Molenberghs et al. 2016). Finally, positive correlations were observed in the striatum and amygdala/hippocampus, as well

as motor cortices. Intriguingly, the seemingly distinct confidence-related network activations at the time of the perceptual decision vs. metacognitive report suggest these regions may encode qualitatively distinct representations of confidence at different times within the trial, for example faster and more automated representations of confidence (see (Lebreton et al. 2015)) around the time of decision, in contrast to metacognitive representations, when explicit evaluation/report are required.

			Peak MNI coordinates (mm)			
Brain region	**BA**	**Laterality**	**X**	**Y**	**Z**	**Z value (peak)**
Positive parametric effect (Z>2.57)						
Striatum (nucleus accumbens / ventral putamen)	-	L	-10	4	-10	4.64
	-	R	12	4	-10	4.09
Lateral orbitofrontal cortex	11/47	L	-28	46	-8	4.46
	47	R	32	38	-6	3.86
Anterior cingulate cortex	32/10	R, L	2	36	6	4.19
Lateral occipital cortex (inferior)	19	L	-42	-68	-10	4.04
	19	R	48	-82	8	3.13
Middle frontal gyrus (anterior)	10	R	40	62	10	3.94
Striatum (dorsal putamen / pallidum)	-	L	-28	-18	2	3.72
Occipital pole	17	R, L	4	-102	8	3.66
Cerebellum	-	R	22	-46	-22	3.55
Inferior temporal gyrus	37	R	54	-46	-16	3.51
Negative parametric effect (Z<-2.57)						
Superior frontal gyrus (supplementary motor area)	6	R	14	12	64	5.62
Dorsomedial prefrontal cortex	6/32	R, L	-6	12	52	4.13
Inferior frontal gyrus	44/45	R	50	16	2	3.95
Precentral gyrus	6	R	50	4	46	3.61
	6	L	-44	2	38	3.46

MNI, Montreal Neurological Institute; L, left hemisphere; R, right hemisphere; BA, approximate Broadmann area.

Table 4.1. Complete list of brain activations correlating with subjects' confidence reports, at the time of stimulus onset (decision phase).

Brain region	BA	Laterality	Peak MNI coordinates (mm) X	Y	Z	Z value (peak)
Positive parametric effect (Z > 2.57)						
Amygdala / Hippocampus	-	R	28	-10	-14	4.16
	-	L	-28	-12	-12	3.27
Putamen	-	L	-22	18	2	4.01
Precentral gyrus	6/4	L	-38	-10	70	3.87
	6	R	38	-14	70	3.04
Negative parametric effect (Z< -2.57)						
Angular gyrus	39	L	-58	-56	34	5.87
Angular gyrus	39	R	60	-54	36	5.82
Superior frontal gyrus / RLPFC	10/9	R	24	58	26	5.84
	10/9	L	-20	52	26	5.2
Inferior frontal gyrus (orbital area) / Anterior insula	13/45	L	-44	24	-8	5.58
	13/45	R	42	22	-6	5.26
Middle frontal gyrus	8/9	R	44	20	42	5.56
	8/9	L	-40	20	42	4.92
Medial frontal gyrus	8/9	L, R	0	42	34	5.19
Inferior frontal gyrus (triangular area)	45	L	-50	22	6	5.02
	45	R	58	30	8	4.94
Precuneus	7	L, R	-2	-68	38	4.51
Occipitotemporal gyrus	37	L	-38	-62	-22	4.34
Posterior cingulate cortex	23	L, R	-2	-26	32	4.76
Middle temporal gyrus (anterior)	20/21	R	50	2	-34	4.36
Thalamus	-	R	10	-10	2	4.35
	-	L	-12	-10	6	3.82
Lingual gyrus	18	L	-2	-80	0	4.14
Calcarine cortex	17	R	16	-90	2	4.14
	17	L	-12	-92	2	3.93
Middle temporal gyrus (posterior)	21/37	R	56	-34	-12	3.93
	21	L	-54	-30	-8	3.82
Inferior occipital gyrus	18	R	28	-90	-10	3.19
Lateral occipital cortex (superior)	19	R	44	-74	20	3.58
	19	L	-40	-88	20	3.43

MNI, Montreal Neurological Institute; L, left hemisphere; R, right hemisphere; BA, approximate Broadmann area; RLPFC, rostrolateral prefrontal cortex

Table 4.2. Complete list of brain activations correlating with subjects' confidence reports, at the time of confidence rating (rating phase).

fMRI correlates of EEG-derived confidence. We used the single-trial variability associated with the confidence discriminating component to construct a parametric EEG-derived fMRI regressor (Y_{CONF} regressor), in order to identify potential brain regions encoding internal representations of early confidence as captured by this EEG component.

Crucial to our approach was modelling the fMRI activation using time-resolved, electrophysiologically-derived signatures of confidence which were specific to each subject. These measures captured the variability in the neural representation of confidence around the perceptual decision itself (i.e., prior to behavioural response), and at a time point of maximum confidence discrimination, thus allowing us to detect its associated spatial correlates with increased temporal and spatial precision, relative to what behavioural ratings and fMRI measurements alone permitted. Importantly, as these signals were only partially correlated with reported confidence, they could potentially provide additional explanatory power in our fMRI model.

This EEG-informed fMRI analysis revealed a large cluster in the ventromedial prefrontal cortex (VMPFC, peak MNI coordinates [-8 40 -14]), extending into the subcallosal region and ventral striatum, and a smaller cluster in the right precentral gyrus (peak MNI coordinates [30 -20 64]), where the BOLD response correlated positively with the EEG-derived confidence discriminating component (Fig. 4.4). Recent studies have linked the VMPFC to confidence in value-based as well as other complex decisions (De Martino et al. 2013, Lebreton et al. 2015), however this region is not typically associated with confidence in perceptual decisions (though see (Heereman et al. 2015)). This finding is consistent with recent work proposing a domain-general role for the VMPFC in encoding confidence (Lebreton et al. 2015), and raises the possibility that this region holds information about early confidence signals emerging prior to the execution of a behavioural choice.

Importantly, we note that the EEG-derived measures, which informed the fMRI analysis, were independent of task difficulty, accuracy, or attention, as discussed in previous sections. Additionally, the GLM model included separate regressors controlling for these variables, and other potential confounds (see

Materials and Methods). In particular, our simultaneous EEG/fMRI approach allowed the introduction of an additional level of control for attentional confounds in the fMRI analysis, namely by including the same EEG-derived index of attention as a nuisance predictor in the GLM model. This regressor showed significant correlation with the intraparietal regions and the frontal eye fields, consistent with the dorsal attentional network thought to be involved in top-down control of visual attention (Corbetta and Shulman 2002).

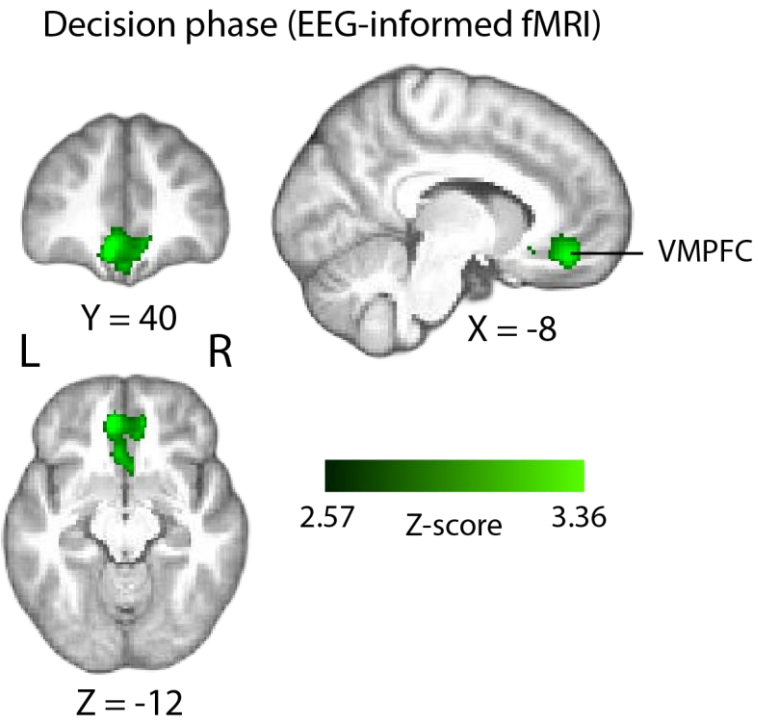

Figure 4.4. EEG-informed fMRI results. Positive parametric modulation of the BOLD signal by EEG-derived single-trial confidence measures (see Materials and Methods), during the decision phase of the trial. Results are reported at $|Z| \geq 2.57$, and cluster-corrected using a resampling procedure (minimum cluster size 162 voxels). *VMPFC*, ventromedial prefrontal cortex.

Next, we asked whether BOLD activation observed in the VMPFC during the perceptual decision period was uniquely associated with the EEG-derived Y_{CONF} regressor, i.e., over and above what could be explained by the behavioural confidence ratings (i.e., the $Ratings_{DEC}$ regressor, Fig. 4.3A) alone. To test this, we compared mean parameter estimates (z-scored beta values) associated with the two predictors, within the VMPFC region identified with the Y_{CONF} regressor.

We found that, across subjects, these were significantly higher for the Y_{CONF} regressor than for the $Ratings_{DEC}$ regressor (paired t-test, $t(23) = 9.48$, $p<.001$). Moreover, VMPFC parameter estimates for the Y_{CONF} regressor remained significantly higher than those associated with the $Ratings_{DEC}$, even when the latter were obtained with a control GLM model that did not include Y_{CONF} as a predictor (paired t-test, $t(23) = 7.99$, $p<.001$). Taken together, these observations indicate that our EEG-derived endogenous measures of confidence were better predictors of VMPFC activity at the time of decision than the post-decision behavioural reports.

The scalp map associated with our confidence discriminating EEG component showed a diffused topography including contributions from several centroparietal electrode sites, however our EEG-derived regressor did not show significant activation in parietal regions. One possibility is that the observed spatial pattern reflects sources of shared variance between the EEG component and confidence ratings themselves (which was otherwise controlled for in our original fMRI analysis). To test this, we ran a separate control GLM analysis where the confidence ratings ($Ratings_{DEC}$) regressor was removed, and found that with this model the Y_{CONF} regressor explained additional variability of the BOLD signal within several regions, including precuneus/PCC regions of the parietal cortex. Notably, activity in these regions has been previously shown to scale with decision confidence (De Martino et al. 2013, White et al. 2014).

Discussion

Here, we used a simultaneous EEG/fMRI approach to investigate the neural correlates of confidence during perceptual decisions. We found that BOLD activation in the VMPFC was uniquely explained by the single-trial variability in an early EEG-derived neural signature of confidence occurring prior to subjects' behavioural expression of response and metacognitive report. Importantly, we showed that this activity surpassed what could be explained by subjects' behavioural reports alone. Our results provide empirical support for the involvement of the VMPFC in confidence of perceptual decisions, consistent with

recent evidence for a domain-general role of the VMPFC in encoding decision confidence. In turn this suggests that the VMPFC may support an early readout of confidence, distinct from explicit metacognitive evaluation.

Our method allowed us to capitalise on the increased explanatory power inherent to our time-resolved internal measures of confidence, to identify relevant activation in the fMRI data. This, in turn, provided a more precise spatiotemporal characterisation than allowed by fMRI measures alone. The observation that the VMPFC holds information about confidence signals occurring prior to behavioural response is intriguing, as it raises novel possibilities for the role of this region in the confidence processing stream. Specifically, the VMPFC may encode early confidence representations (at, or near, the time of decision), which in turn could have important adaptive functions in influencing action that follows from the perceptual decision, and potentially informing the choice (Lak et al. 2017). Additionally, such signals may be qualitatively different from confidence estimates available at the time of report as the latter are likely to undergo additional processing that continues after a choice is made (Moran et al. 2015, Resulaj et al. 2009, Pleskac and Busemeyer 2010).

Computational and neurobiological accounts of confidence processing have proposed architectures by which a first-level form of confidence in a decision emerges as a natural property of the neural processes supporting the decision, which in turn is read out (i.e., summarised) by separate higher-order monitoring network(s) (Pouget et al. 2016, Insabato et al. 2010, Meyniel et al. 2015). As the VMPFC is not typically known to support perceptual decision processes, the VMPFC confidence signals we observe here are thus likely to represent a readout of confidence-related information from upstream regions.

Consistent with a role as a monitoring module providing a confidence readout, recent work suggests the VMPFC may encode confidence in a task-independent and possibly domain-general manner. Specifically, several functional neuroimaging studies have shown positive modulation of VMPFC activation by confidence, across a range of decision making tasks (Lebreton et al. 2015, De Martino et al. 2013, Heereman et al. 2015, Rolls et al. 2010a). Notably, one study showed that fMRI activation in the VMPFC was modulated by confidence

across four different tasks involving both value-based and non-value based rating judgments (Lebreton et al. 2015). Furthermore, evidence from memory-related decision making research appears to also implicate the VMPFC in confidence processing (see Hebscher and Gilboa, 2016, for a review). Our results in the present study complement current literature by bringing empirical support for the involvement of VMPFC in perceptual decision making.

The observation that the VMPFC, a region known for its involvement in choice-related subjective valuation (Rangel and Hare 2010, Bartra et al. 2013, Philiastides et al. 2010, Pisauro et al. 2017) encodes confidence signals during perceptual decisions raises an interesting possibility for interpreting our results. Our behavioural paradigm did not involve an explicit reward/feedback manipulation and accordingly, the observed confidence-related activation cannot be interpreted as an externally driven value signal. Instead, as has been suggested previously (Barron et al. 2015, Lebreton et al. 2015), a likely explanation is that, by being an internal measure of performance accuracy, confidence is inherently valuable. Such a signal may represent *implicit* reward and possibly act as a teaching signal (Lak et al. 2017, Guggenmos et al. 2016, Daniel and Pollmann 2012) to drive learning (e.g., perceptual learning (Diaz et al. 2017, Kahnt et al. 2011, Law and Gold 2009).

In line with this interpretation, Hebart et al. (2016) observed positive correlation with confidence in the ventral striatum, a region known for its involvement in reward (O'Doherty et al. 2004). Authors suggest that confidence signals in this region may play a role in confidence-driven learning, such that feelings of reward associated with a choice reinforce optimal behavior on subsequent choices. A different study (Guggenmos et al. 2016) demonstrated that regions of the human mesolimbic dopamine system, namely the striatum and ventral tegmental area, encoded both anticipation and prediction error related to decision confidence (i.e., in the absence of feedback), similar to what is typically observed during reinforcement learning tasks where feedback is explicit (Fouragnan et al. 2015, Preuschoff et al. 2006, Fouragnan et al. 2017). Importantly, these effects were predictive of subjects' perceptual learning efficiency. Thus, confidence in valuation/reward networks could be propagated

back to the decision systems to optimize the dynamics of the decision process, for example by means of a reinforcement-learning mechanism.

In conclusion, we showed that by employing a simultaneous EEG/fMRI approach, we were able to localise an early representation of confidence in the brain with higher spatiotemporal precision than allowed by fMRI alone. In doing so, we provided novel empirical evidence for the encoding of a generalised confidence readout signal in the VMPFC preceding explicit metacognitive report. Our findings provide a starting point for further investigations into the neural dynamics of confidence formation in the human brain and its interaction with other cognitive processes such as learning, and the choice itself.

Chapter 5. General Discussion

Overview

The sense of confidence in our judgments is a vital factor in our interactions with the environment. From decisions as complex as choosing a career, to as simple as discerning between objects in dim light, the degree to which we believe we are correct influences our actions and subsequent decisions. As discussed in the introductory chapter, there has been significant progress towards uncovering the neural basis of confidence-related processes within the past decade. Combined work in animals and humans suggests that even for simple perceptual decisions, the construction of confidence involves the interaction of multiple neural networks (Grimaldi et al. 2015, Meyniel et al. 2015). Regions of the prefrontal cortex have been implicated in higher-order monitoring processes associated with metacognitive appraisal (Fleming et al. 2012, Fleming et al. 2010), however confidence-related information has also been detected at earlier stages of processing, as early as the decision process itself (Zylberberg et al. 2016, Kiani and Shadlen 2009, Middlebrooks and Sommer 2012, Zizlsperger et al. 2014). Human studies investigating the neural correlates of confidence have focused primarily on identifying neural signals that support metacognitive (i.e., monitoring) processes, however, to begin to understand the complex neural dynamics involved in confidence processing, it is relevant to identify neural representations that may precede or contribute to these higher-order signals. To this end, the current thesis sought to offer a more complete characterisation of confidence representations occurring near the time of the perceptual decision, in the human brain.

Chapters 2 and 3 relied on the high temporal precision of EEG measurements to investigate the neural mechanisms supporting confidence formation during a face vs. car categorisation task. More specifically, in Chapter 2 we asked whether, as suggested by some theoretical accounts (Pouget et al. 2016, Meyniel et al. 2015, Insabato et al. 2010) and animal neurophysiology (Kiani and Shadlen

2009), confidence may be represented in the neural activity that underlies the decision process. Chapter 3 further investigated whether rhythmic activity in these time-resolved signals may contain additional information about the neural mechanisms supporting confidence processes. Finally, building on findings from our first study (Chapter 2), and capitalising on single-trial neural signatures of confidence estimated with EEG, in Chapter 4 we recorded simultaneous EEG/fMRI measurements aiming to identify potential networks linked with this activity. We hypothesised that the trial-to-trial variability within these endogenous measures would capture early confidence-related signals (i.e., occurring prior to overt commitment to choice or explicit metacognitive evaluation) with higher accuracy than allowed with metacognitive reports only (which may reflect additional influences resulting from post-decisional processes, Pleskac and Busemeyer, 2010, Moran et al., 2015, Yu et al., 2015, van den Berg et al., 2016a).

Key findings

Using a single-trial multivariate analysis of the EEG, the first study revealed confidence-discriminating activity peaking on average 600 ms after stimulus onset. This neural representation of confidence appeared to be reflected in the rate of evidence accumulation that characterised the perceptual decision, supporting the idea of a shared mechanisms underlying confidence and choice (Kiani and Shadlen 2009). The scalp topography associated with this activity showed centroparietal electrode contributions, similar to neural representations of evidence accumulation that have been identified across different tasks and sensory modalities (Kelly and O'Connell 2013, O'Connell et al. 2012, Philiastides et al. 2014). Complementing this observation, the second study revealed that a separate, motor-preparatory signal (Pfurtscheller and Lopes da Silva 1999), also carried information about subjects' confidence. Specifically, oscillatory activity in the alpha- and beta-bands (approx. 10-22 Hz) over the contralateral hemisphere relative to the motor effector was reduced during high- vs. low-confidence choices. This effect began shortly after stimulus onset (~300 ms) and peaked around 600-650 ms (i.e., seconds before a motor response), showing

large overlap with the neural representation of confidence in the time-domain (though it must be noted that oscillatory activity estimates are inherently less temporally precise than time-domain measurements). While distinct from supramodal decision signals (O'Connell et al. 2012), motor-preparatory signals have been shown to reflect the evolving decision process (O'Connell et al. 2012, Donner et al. 2009, de Lange et al. 2013), consistent with a continuous flow of information from regions that encode the decision (Selen et al. 2012, Gold and Shadlen 2003). It follows that if such signals hold information about the decision formation, they may also carry information about confidence inherent to this process. Together, our EEG data indicate that neural representations of confidence may simultaneously be available within neural circuits relevant to the decision and the impending action. The view that confidence and choice are encoded within the same neural code is also consistent with Bayesian accounts of neural processing (Knill and Pouget 2004) postulating that perceptual choices are represented as probability distributions, with confidence thus being reflected in the neural code that represents the decision (Meyniel et al. 2015, Pouget et al. 2016). This of course does not exclude the possibility that confidence information, as revealed by our EEG data, may be read out and integrated by higher-order structures (De Martino et al. 2013, Insabato et al. 2010, Hebart et al. 2016, Fleming et al. 2012).

In Chapter 4, we recorded simultaneous EEG and fMRI measurements while subjects performed a random-dot motion discrimination task. Single-trial analysis of the EEG revealed a confidence-discriminating component whose temporal profile and scalp topography matched our results from the face vs. car categorisation task in Chapter 2, pointing to a non-task-specific neural representation of confidence. Importantly, our EEG-informed fMRI analysis showed that single-trial variability within the EEG-derived neural signatures of confidence uniquely explained activation in the ventromedial prefrontal cortex (VMPFC) during the decision phase of the trial. As discussed in Chapter 4, this region has been shown to encode confidence in several decision making tasks (Lebreton et al. 2015, De Martino et al. 2013, Heereman et al. 2015), however its role in perceptual confidence is not clear. As the VMPFC is not typically

associated with perceptual decision making, we speculated that activity here could represent a higher-order readout of confidence-related information. Multiple regions have been shown to encode confidence independently of the perceptual decision, including the orbitofrontal cortex in the rat brain (Lak et al. 2014), pulvinar in the monkey (Komura et al. 2013), and the PFC (Fleming et al. 2012, Fleming et al. 2010, Rounis et al. 2010, Lau and Passingham 2006, Baird et al. 2013) and striatum (Hebart et al. 2016) in humans, suggesting confidence may be read out and used by multiple neural circuits, potentially serving different functional purposes.

Importantly, our results implied that representations of confidence in the VMPFC are better explained by early internal confidence signatures (extracted prior to overt choice or explicit metacognitive evaluation) than by behavioural confidence reports, which are theorised to rely on additional noisy post-decisional processing (Pleskac and Busemeyer 2010, Moran et al. 2015, Yu et al. 2015). One explanation for this finding could be, for example, shorter post-decisional processing delays that introduce additional changes to the confidence readout. Potentially in line with this is the observation that the VMPFC seems to support an automatic readout of confidence (i.e., in the absence of explicit report) (Lebreton et al. 2015). Similarly, the vmPFC is thought to encode an early and automatic "feeling of rightness" (Moscovitch and Winocur 2002, Hebscher and Gilboa 2016) in memory judgments. As such, whereas explicit metacognitive evaluation may involve additional post-decisional processing, and/or integration of information from multiple sources (e.g., action- or choice-related information, Fleming et al., 2015), regions such as the VMPFC may encode an automatic (potentially faster) confidence readout. Such a distinction could be made for example between the VMPFC and the anterior PFC. The RLPFC appears consistent with a role in deliberate metacognitive evaluation (i.e., explicit engagement in self-monitoring). Namely, haemodynamic responses in this area scale with confidence during explicit metacognitive report, as shown in our study (Chapter 4) and other experiments (Fleming et al. 2012, Hilgenstock et al. 2014), and are more pronounced during confidence report than during a

control task (Fleming et al. 2012). Follow up studies could explicitly investigate this potential distinction.

Limitations and future directions

It is worth noting that while the simultaneous EEG/fMRI approach offers clear advantages over the use of these two techniques in isolation, it is nevertheless not free of limitations. In particular, our approach relied on using early EEG-derived neural signatures of confidence to spatially identify networks that might be functionally linked to these early signals. However, as EEG recorded at the scalp surface likely contains mixed inputs from multiple structures, confidence-related contributions are difficult to entirely separate from other sources of variability, and thus these signals may still contain influences unrelated to confidence on a trial-by-trial basis. While our interpretations of the VMPFC activation pattern appears consistent with existing literature, it will be important to obtain additional validation for the role of this region from follow-up studies.

The advantages of our approach nevertheless seem to outweigh its limitations. Our results can be used to create novel data-driven hypotheses which in turn can inform future experiments. Here, we identified the VMPFC as a candidate region for processing confidence in perceptual decision making near the time of the decision. Future studies can further evaluate its functional role, as well as its causal contributions to behaviour. In particular, it will be important to formally establish whether the VMPFC plays a causal role in metacognitive evaluation for perceptual decisions, i.e., whether disrupting activity in this region may affect confidence independently of performance, as has been observed in other regions of the animal and human brain (Rounis et al. 2010, Komura et al. 2013, Lak et al. 2014). Additionally, given the known role of this region in subjective value processing (Philiastides et al. 2010, Rangel and Hare 2010), future studies can explicitly investigate whether confidence-related responses in the VMPFC may potentially play a role in learning in the absence of feedback (Daniel and

Pollmann 2012, Guggenmos et al. 2016, Lak et al. 2017) by acting as implicit reward/valuation signals (Lebreton et al. 2015, Barron et al. 2015).

Together, empirical findings from our three studies reinforce the observation that information about perceptual confidence is represented across multiple neural systems. Additionally, we provide novel insights into the underlying neural mechanisms and spatiotemporal representations of confidence in the human brain, as well as demonstrate how the simultaneous EEG/fMRI approach can be used to characterise these. Modelling approaches may offer additional insights with respect to the computations by which the observed neural signatures of confidence are generated. Variants of sequential-sampling-type models have been used for determining confidence in animals and humans at the time of choice (Kepecs et al. 2008, De Martino et al. 2013, Vickers 1979, Kiani and Shadlen 2009), and thus could serve as basis for extensions of this work.

Conclusion

The current thesis presented empirical findings from three studies which investigated the neural correlates of confidence during human perceptual decision making. We demonstrated that, similarly to observations from animal literature, confidence-related information could be inferred from the same process of evidence accumulation supporting decision formation. Interestingly, confidence-discriminating neural activity was simultaneously present in motor-preparatory signals, consistent with a continuous flow of decision- and confidence-related information into the sensorimotor systems. Finally, we showed that activation in the VMPFC explained variability in internal confidence representations identified near the time of the decision (i.e., prior to subjects' overt response or explicit metacognitive evaluation), in line with a role of this region in encoding early and/or automatic representation of confidence. Our results represent a step towards a more complete characterisation of the neural dynamics involved in confidence processing, and provide a tool for continuing to disentangle the neural sources contributing to human confidence and metacognition.

References

Alegre, M., Labarga, A., Gurtubay, I. G., Iriarte, J., Malanda, A. and Artieda, J. (2003) 'Movement-related changes in cortical oscillatory activity in ballistic, sustained and negative movements', *Exp Brain Res,* 148(1), 17-25.

Allen, M., Glen, J. C., Mullensiefen, D., Schwarzkopf, D. S., Fardo, F., Frank, D., Callaghan, M. F. and Rees, G. (2017) 'Metacognitive ability correlates with hippocampal and prefrontal microstructure', *Neuroimage,* 149, 415-423.

Baird, B., Smallwood, J., Gorgolewski, K. J. and Margulies, D. S. (2013) 'Medial and lateral networks in anterior prefrontal cortex support metacognitive ability for memory and perception', *J Neurosci,* 33(42), 16657-65.

Baranski, J. V. and Petrusic, W. M. (1994) 'The calibration and resolution of confidence in perceptual judgments', *Percept Psychophys,* 55(4), 412-28.

Baranski, J. V. and Petrusic, W. M. (1998) 'Probing the locus of confidence judgments: experiments on the time to determine confidence', *J Exp Psychol Hum Percept Perform,* 24(3), 929-45.

Baranski, J. V. and Petrusic, W. M. (1999) 'Realism of confidence in sensory discrimination', *Percept Psychophys,* 61(7), 1369-83.

Barron, H. C., Garvert, M. M. and Behrens, T. E. (2015) 'Reassessing VMPFC: full of confidence?', *Nat Neurosci,* 18(8), 1064-6.

Bartra, O., McGuire, J. T. and Kable, J. W. (2013) 'The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value', *Neuroimage,* 76, 412-27.

Baumgarten, T. J., Schnitzler, A. and Lange, J. (2016) 'Prestimulus Alpha Power Influences Tactile Temporal Perceptual Discrimination and Confidence in Decisions', *Cereb Cortex,* 26(3), 891-903.

Blank, H., Biele, G., Heekeren, H. R. and Philiastides, M. G. (2013) 'Temporal characteristics of the influence of punishment on perceptual decision making in the human brain', *J Neurosci,* 33(9), 3939-52.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P. and Cohen, J. D. (2006) 'The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks', *Psychol Rev,* 113(4), 700-65.

Boldt, A. and Yeung, N. (2015) 'Shared neural markers of decision confidence and error detection', *J Neurosci,* 35(8), 3478-84.

Cheadle, S., Wyart, V., Tsetsos, K., Myers, N., de Gardelle, V., Herce Castanon, S. and Summerfield, C. (2014) 'Adaptive Gain Control during Human Perceptual Choice', *Neuron,* 81(6), 1429-41.

Chen, R., Yaseen, Z., Cohen, L. G. and Hallett, M. (1998) 'Time course of corticospinal excitability in reaction time and self-paced movements', *Ann Neurol,* 44(3), 317-25.

Cheyne, D. O. (2013) 'MEG studies of sensorimotor rhythms: a review', *Exp Neurol,* 245, 27-39.

Cohen, M. X. (2014) *Analyzing neural time series data: theory and practice,* MIT Press.

Cohen, M. X. (2016) 'Midfrontal theta tracks action monitoring over multiple interactive time scales', *Neuroimage,* 141, 262-72.

Corbetta, M. and Shulman, G. L. (2002) 'Control of goal-directed and stimulus-driven attention in the brain', *Nat Rev Neurosci,* 3(3), 201-15.

Crone, N. E., Sinai, A. and Korzeniewska, A. (2006) 'High-frequency gamma oscillations and human brain mapping with electrocorticography', *Prog Brain Res,* 159, 275-95.

Dakin, S. C., Hess, R. F., Ledgeway, T. and Achtman, R. L. (2002) 'What causes non-monotonic tuning of fMRI response to noisy images?', *Curr Biol,* 12(14), R476-7; author reply R478.

Daniel, R. and Pollmann, S. (2012) 'Striatal activations signal prediction errors on confidence in the absence of external feedback', *Neuroimage,* 59(4), 3457-67.

de Lafuente, V., Jazayeri, M. and Shadlen, M. N. (2015) 'Representation of accumulating evidence for a decision in two parietal areas', *J Neurosci,* 35(10), 4306-18.

De Lafuente, V. and Romo, R. (2011) 'Dopamine neurons code subjective sensory experience and uncertainty of perceptual decisions', *Proc Natl Acad Sci U S A,* 108(49), 19767-71.

de Lange, F. P., Jensen, O. and Dehaene, S. (2010) 'Accumulation of evidence during sequential decision making: the importance of top-down factors', *J Neurosci,* 30(2), 731-8.

de Lange, F. P., Rahnev, D. A., Donner, T. H. and Lau, H. (2013) 'Prestimulus oscillatory activity over motor cortex reflects perceptual expectations', *J Neurosci,* 33(4), 1400-10.

De Martino, B., Fleming, S. M., Garrett, N. and Dolan, R. J. (2013) 'Confidence in value-based choice', *Nat Neurosci,* 16(1), 105-10.

Debettencourt, M., Goldman, R., Brown, T. and Sajda, P. (2011) 'Adaptive Thresholding for Improving Sensitivity in Single-Trial Simultaneous EEG/fMRI', *Front Psychol,* 2, 91.

Diaz, J. A., Queirazza, F. and Philiastides, M. G. (2017) 'Perceptual learning alters post-sensory processing in human decision making', *Nature Human Behaviour,* 1.

Ding, L. and Gold, J. I. (2013) 'The basal ganglia's contributions to perceptual decision making', *Neuron,* 79(4), 640-9.

Donner, T. H. and Siegel, M. (2011) 'A framework for local cortical oscillation patterns', *Trends Cogn Sci,* 15(5), 191-9.

Donner, T. H., Siegel, M., Fries, P. and Engel, A. K. (2009) 'Buildup of choice-predictive activity in human motor cortex during perceptual decision making', *Curr Biol,* 19(18), 1581-5.

Donner, T. H., Siegel, M., Oostenveld, R., Fries, P., Bauer, M. and Engel, A. K. (2007) 'Population activity in the human dorsal pathway predicts the accuracy of visual motion detection', *J Neurophysiol,* 98(1), 345-59.

Dreher, J. C., Kohn, P. and Berman, K. F. (2006) 'Neural coding of distinct statistical properties of reward information in humans', *Cereb Cortex,* 16(4), 561-73.

Duda, R. O., Hart, P. E. and Stork, D. G. (2001) 'Pattern classification', *New York: John Wiley, Section,* 10, l.

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(4), 291.

Fetsch, C. R., Kiani, R., Newsome, W. T. and Shadlen, M. N. (2014) 'Effects of cortical microstimulation on confidence in a perceptual decision', *Neuron,* 83(4), 797-804.

Filimon, F., Philiastides, M. G., Nelson, J. D., Kloosterman, N. A. and Heekeren, H. R. (2013) 'How embodied is perceptual decision making? Evidence for separate processing of perceptual and motor decisions', *J Neurosci,* 33(5), 2121-36.

Fleck, M. S., Daselaar, S. M., Dobbins, I. G. and Cabeza, R. (2006) 'Role of prefrontal and anterior cingulate regions in decision-making processes shared by memory and nonmemory tasks', *Cereb Cortex,* 16(11), 1623-30.

Fleming, S. M. and Daw, N. D. (2017) 'Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation', *Psychol Rev,* 124(1), 91-114.

Fleming, S. M. and Dolan, R. J. (2010) 'Effects of loss aversion on post-decision wagering: implications for measures of awareness', *Conscious Cogn,* 19(1), 352-63.

Fleming, S. M. and Dolan, R. J. (2012) 'The neural basis of metacognitive ability', *Philos Trans R Soc Lond B Biol Sci,* 367(1594), 1338-49.

Fleming, S. M., Huijgen, J. and Dolan, R. J. (2012) 'Prefrontal contributions to metacognition in perceptual decision making', *J Neurosci,* 32(18), 6117-25.

Fleming, S. M., Maniscalco, B., Ko, Y., Amendi, N., Ro, T. and Lau, H. (2015) 'Action-specific disruption of perceptual confidence', *Psychol Sci,* 26(1), 89-98.

Fleming, S. M., Ryu, J., Golfinos, J. G. and Blackmon, K. E. (2014) 'Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions', *Brain,* 137(Pt 10), 2811-22.

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J. and Rees, G. (2010) 'Relating introspective accuracy to individual differences in brain structure', *Science,* 329(5998), 1541-3.

Folke, T., Jacobsen, C., Fleming, S. M. and De Martino, B. (2016) 'Explicit representation of confidence informs future value-based decisions', *Nature Human Behaviour,* 1.

Forstmann, B. U., Dutilh, G., Brown, S., Neumann, J., von Cramon, D. Y., Ridderinkhof, K. R. and Wagenmakers, E. J. (2008) 'Striatum and pre-SMA facilitate decision-making under time pressure', *Proc Natl Acad Sci U S A,* 105(45), 17538-42.

Fouragnan, E., Queirazza, F., Retzler, C., Mullinger, K. J. and Philiastides, M. G. (2017) 'Spatiotemporal neural characterization of prediction error valence and surprise during reward learning in humans', *Sci Rep,* 7(1), 4762.

Fouragnan, E., Retzler, C., Mullinger, K. and Philiastides, M. G. (2015) 'Two spatiotemporally distinct value systems shape reward-based learning in the human brain', *Nat Commun,* 6, 8107.

Friston, K., Harrison, L., Daunizeau, J., Kiebel, S., Phillips, C., Trujillo-Barreto, N., Henson, R., Flandin, G. and Mattout, J. (2008) 'Multiple sparse priors for the M/EEG inverse problem', *Neuroimage,* 39(3), 1104-20.

Gherman, S. and Philiastides, M. G. (2015) 'Neural representations of confidence emerge from the process of decision formation during perceptual choices', *Neuroimage,* 106, 134-43.

Gold, J. I. and Shadlen, M. N. (2000) 'Representation of a perceptual decision in developing oculomotor commands', *Nature,* 404(6776), 390-4.

Gold, J. I. and Shadlen, M. N. (2003) 'The influence of behavioral context on the representation of a perceptual decision in developing oculomotor commands', *J Neurosci,* 23(2), 632-51.

Gold, J. I. and Shadlen, M. N. (2007) 'The neural basis of decision making', *Annu Rev Neurosci,* 30, 535-74.

Goldman, R. I., Wei, C. Y., Philiastides, M. G., Gerson, A. D., Friedman, D., Brown, T. R. and Sajda, P. (2009) 'Single-trial discrimination for integrating simultaneous EEG and fMRI: identifying cortical areas contributing to trial-to-trial variability in the auditory oddball task', *Neuroimage,* 47(1), 136-47.

Grabenhorst, F. and Rolls, E. T. (2011) 'Value, pleasure and choice in the ventral prefrontal cortex', *Trends Cogn Sci,* 15(2), 56-67.

Greve, D. N. and Fischl, B. (2009) 'Accurate and robust brain image alignment using boundary-based registration', *Neuroimage,* 48(1), 63-72.

Grimaldi, P., Lau, H. and 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', *Neurosci Biobehav Rev,* 55, 88-97.

Grinband, J., Hirsch, J. and Ferrera, V. P. (2006) 'A neural representation of categorization uncertainty in the human brain', *Neuron,* 49(5), 757-63.

Guggenmos, M., Wilbertz, G., Hebart, M. N. and Sterzer, P. (2016) 'Mesolimbic confidence signals guide perceptual learning in the absence of external feedback', *Elife,* 5.

Haegens, S., Nacher, V., Hernandez, A., Luna, R., Jensen, O. and Romo, R. (2011) 'Beta oscillations in the monkey sensorimotor network reflect somatosensory decision making', *Proc Natl Acad Sci U S A,* 108(26), 10708-13.

Hebart, M. N., Schriever, Y., Donner, T. H. and Haynes, J. D. (2014) 'The Relationship between Perceptual Decision Variables and Confidence in the Human Brain', *Cereb Cortex*.

Hebart, M. N., Schriever, Y., Donner, T. H. and Haynes, J. D. (2016) 'The Relationship between Perceptual Decision Variables and Confidence in the Human Brain', *Cereb Cortex,* 26(1), 118-30.

Hebscher, M. and Gilboa, A. (2016) 'A boost of confidence: The role of the ventromedial prefrontal cortex in memory, decision-making, and schemas', *Neuropsychologia,* 90, 46-58.

Heekeren, H. R., Marrett, S., Bandettini, P. A. and Ungerleider, L. G. (2004) 'A general mechanism for perceptual decision-making in the human brain', *Nature,* 431(7010), 859-62.

Heekeren, H. R., Marrett, S., Ruff, D. A., Bandettini, P. A. and Ungerleider, L. G. (2006) 'Involvement of human left dorsolateral prefrontal cortex in perceptual decision making is independent of response modality', *Proc Natl Acad Sci U S A,* 103(26), 10023-8.

Heekeren, H. R., Marrett, S. and Ungerleider, L. G. (2008) 'The neural systems that mediate human perceptual decision making', *Nat Rev Neurosci,* 9(6), 467-79.

Heereman, J., Walter, H. and Heekeren, H. R. (2015) 'A task-independent neural representation of subjective certainty in visual perception', *Front Hum Neurosci,* 9, 551.

Hilgenstock, R., Weiss, T. and Witte, O. W. (2014) 'You'd better think twice: post-decision perceptual confidence', *Neuroimage,* 99, 323-31.

Ho, T. C., Brown, S. and Serences, J. T. (2009) 'Domain general mechanisms of perceptual decision making in human cortex', *J Neurosci,* 29(27), 8675-87.

Horwitz, G. D. and Newsome, W. T. (1999) 'Separate signals for target selection and movement specification in the superior colliculus', *Science,* 284(5417), 1158-61.

Huk, A. C. and Shadlen, M. N. (2005) 'Neural activity in macaque parietal cortex reflects temporal integration of visual motion signals during perceptual decision making', *J Neurosci,* 25(45), 10420-36.

Insabato, A., Pannunzi, M., Rolls, E. T. and Deco, G. (2010) 'Confidence-related decision making', *J Neurophysiol,* 104(1), 539-47.

Jeffreys, D. A. (1996) 'Evoked potential studies of face and object processing', *Visual Cognition,* 3(1), 1-38.

Jordan, M. I. and Jacobs, R. A. (1994) 'Hierarchical Mixtures of Experts and the Em Algorithm', *Neural Computation,* 6(2), 181-214.

Kable, J. W. and Glimcher, P. W. (2007) 'The neural correlates of subjective value during intertemporal choice', *Nat Neurosci,* 10(12), 1625-33.

Kahnt, T., Grueschow, M., Speck, O. and Haynes, J. D. (2011) 'Perceptual learning and decision-making in human medial frontal cortex', *Neuron,* 70(3), 549-59.

Kelly, S. P. and O'Connell, R. G. (2013) 'Internal and external influences on the rate of sensory evidence accumulation in the human brain', *J Neurosci,* 33(50), 19434-41.

Kepecs, A. and Mainen, Z. F. (2012) 'A computational framework for the study of confidence in humans and animals', *Philos Trans R Soc Lond B Biol Sci,* 367(1594), 1322-37.

Kepecs, A., Uchida, N., Zariwala, H. A. and Mainen, Z. F. (2008) 'Neural correlates, computation and behavioural impact of decision confidence', *Nature,* 455(7210), 227-31.

Kiani, R. and Shadlen, M. N. (2009) 'Representation of confidence associated with a decision by neurons in the parietal cortex', *Science,* 324(5928), 759-64.

Kim, J. N. and Shadlen, M. N. (1999) 'Neural correlates of a decision in the dorsolateral prefrontal cortex of the macaque', *Nat Neurosci,* 2(2), 176-85.

Knill, D. C. and Pouget, A. (2004) 'The Bayesian brain: the role of uncertainty in neural coding and computation', *Trends Neurosci,* 27(12), 712-9.

Knutson, B., Taylor, J., Kaufman, M., Peterson, R. and Glover, G. (2005) 'Distributed neural representation of expected value', *J Neurosci,* 25(19), 4806-12.

Komura, Y., Nikkuni, A., Hirashima, N., Uetake, T. and Miyamoto, A. (2013) 'Responses of pulvinar neurons reflect a subject's confidence in visual categorization', *Nat Neurosci,* 16(6), 749-55.

Krueger, P. M., van Vugt, M. K., Simen, P., Nystrom, L., Holmes, P. and Cohen, J. D. (2017) 'Evidence accumulation detected in BOLD signal using slow perceptual decision making', *J Neurosci Methods,* 281, 21-32.

Kubanek, J., Hill, N. J., Snyder, L. H. and Schalk, G. (2015) 'Cortical alpha activity predicts the confidence in an impending action', *Front Neurosci,* 9, 243.

Lak, A., Costa, G. M., Romberg, E., Koulakov, A. A., Mainen, Z. F. and Kepecs, A. (2014) 'Orbitofrontal cortex is required for optimal waiting based on decision confidence', *Neuron,* 84(1), 190-201.

Lak, A., Nomoto, K., Keramati, M., Sakagami, M. and Kepecs, A. (2017) 'Midbrain Dopamine Neurons Signal Belief in Choice Accuracy during a Perceptual Decision', *Curr Biol,* 27(6), 821-832.

Lau, H. C. and Passingham, R. E. (2006) 'Relative blindsight in normal observers and the neural correlate of visual consciousness', *Proc Natl Acad Sci U S A,* 103(49), 18763-8.

Law, C. T. and Gold, J. I. (2009) 'Reinforcement learning can account for associative and perceptual learning on a visual-decision task', *Nat Neurosci,* 12(5), 655-63.

Lebreton, M., Abitbol, R., Daunizeau, J. and Pessiglione, M. (2015) 'Automatic integration of confidence in the brain valuation signal', *Nat Neurosci,* 18(8), 1159-67.

Liu, J., Higuchi, M., Marantz, A. and Kanwisher, N. (2000) 'The selectivity of the occipitotemporal M170 for faces', *Neuroreport,* 11(2), 337-341.

Liu, T. and Pleskac, T. J. (2011) 'Neural correlates of evidence accumulation in a perceptual decision task', *J Neurophysiol,* 106(5), 2383-98.

Ma, W. J., Beck, J. M., Latham, P. E. and Pouget, A. (2006) 'Bayesian inference with probabilistic population codes', *Nat Neurosci,* 9(11), 1432-8.

Maniscalco, B. and Lau, H. (2016) 'The signal processing architecture underlying subjective reports of sensory awareness', *Neurosci Conscious,* 2016(1).

Maris, E. and Oostenveld, R. (2007) 'Nonparametric statistical testing of EEG- and MEG-data', *J Neurosci Methods,* 164(1), 177-90.

Mazaheri, A. and Jensen, O. (2006) 'Posterior alpha activity is not phase-reset by visual stimuli', *Proc Natl Acad Sci U S A,* 103(8), 2948-52.

Mazurek, M. E., Roitman, J. D., Ditterich, J. and Shadlen, M. N. (2003) 'A role for neural integrators in perceptual decision making', *Cereb Cortex,* 13(11), 1257-69.

McFarland, D. J., Miner, L. A., Vaughan, T. M. and Wolpaw, J. R. (2000) 'Mu and beta rhythm topographies during motor imagery and actual movements', *Brain Topogr,* 12(3), 177-86.

Meyniel, F., Sigman, M. and Mainen, Z. F. (2015) 'Confidence as Bayesian Probability: From Neural Origins to Behavior', *Neuron,* 88(1), 78-92.

Middlebrooks, P. G. and Sommer, M. A. (2012) 'Neuronal correlates of metacognition in primate frontal cortex', *Neuron,* 75(3), 517-30.

Mitra, P. P. and Pesaran, B. (1999) 'Analysis of dynamic brain imaging data', *Biophys J,* 76(2), 691-708.

Molenberghs, P., Trautwein, F. M., Bockler, A., Singer, T. and Kanske, P. (2016) 'Neural correlates of metacognitive ability and of feeling confident: a large-scale fMRI study', *Soc Cogn Affect Neurosci*.

Moran, R., Teodorescu, A. R. and Usher, M. (2015) 'Post choice information integration as a causal determinant of confidence: Novel data and a computational account', *Cogn Psychol,* 78, 99-147.

Moscovitch, M. and Winocur, G. (2002) 'The frontal cortex and working with memory', *Principles of frontal lobe function*, 188-209.

Mulder, M. J., van Maanen, L. and Forstmann, B. U. (2014) 'Perceptual decision neurosciences - a model-based review', *Neuroscience,* 277, 872-84.

Murphy, P. R., Robertson, I. H., Harty, S. and O'Connell, R. G. (2015) 'Neural evidence accumulation persists after choice to inform metacognitive judgments', *Elife,* 4.

Navajas, J., Bahrami, B. and Latham, P. E. (2016) 'Post-decisional accounts of biases in confidence', *Curr Opin Behav Sci,* 11, 55-60.

Neuper, C. and Pfurtscheller, G. (2001) 'Event-related dynamics of cortical rhythms: frequency-specific features and functional correlates', *Int J Psychophysiol,* 43(1), 41-58.

Newsome, W. T., Britten, K. H. and Movshon, J. A. (1989) 'Neuronal correlates of a perceptual decision', *Nature,* 341(6237), 52-4.

Newsome, W. T. and Pare, E. B. (1988) 'A selective impairment of motion perception following lesions of the middle temporal visual area (MT)', *The Journal of Neuroscience,* 8(6), 2201-2211.

Newsome, W. T. and Pare, E. B. (1988) 'A selective impairment of motion perception following lesions of the middle temporal visual area (MT)', *J Neurosci,* 8(6), 2201-11.

O'Connell, R. G., Dockree, P. M. and Kelly, S. P. (2012) 'A supramodal accumulation-to-bound signal that determines perceptual decisions in humans', *Nat Neurosci,* 15(12), 1729-35.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K. and Dolan, R. J. (2004) 'Dissociable roles of ventral and dorsal striatum in instrumental conditioning', *Science,* 304(5669), 452-4.

O'Doherty, J. P. (2004) 'Reward representations and reward-related learning in the human brain: insights from neuroimaging', *Curr Opin Neurobiol,* 14(6), 769-76.

Odegaard, B., Grimaldi, P., Cho, S. H., Peters, M. A. K., Lau, H. and Basso, M. A. (2017) 'Superior Colliculus Neuronal Ensemble Activity Signals Optimal Rather Than Subjective Confidence', *bioRxiv*.

Oostenveld, R., Fries, P., Maris, E. and Schoffelen, J. M. (2011) 'FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data', *Comput Intell Neurosci,* 2011, 156869.

Parra, L., Alvino, C., Tang, A., Pearlmutter, B., Yeung, N., Osman, A. and Sajda, P. (2002) 'Linear spatial integration for single-trial detection in encephalography', *Neuroimage,* 17(1), 223-30.

Parra, L. C., Spence, C. D., Gerson, A. D. and Sajda, P. (2005) 'Recipes for the linear analysis of EEG', *Neuroimage,* 28(2), 326-41.

Paul, E. J., Smith, J. D., Valentin, V. V., Turner, B. O., Barbey, A. K. and Ashby, F. G. (2015) 'Neural networks underlying the metacognitive uncertainty response', *Cortex,* 71, 306-22.

Peirce, C. S. and Jastrow, J. (1884) 'On small differences in sensation'.

Peirce, J. W. (2007) 'PsychoPy--Psychophysics software in Python', *J Neurosci Methods,* 162(1-2), 8-13.

Persaud, N., McLeod, P. and Cowey, A. (2007) 'Post-decision wagering objectively measures awareness', *Nat Neurosci,* 10(2), 257-61.

Peters, M. A., Thesen, T., Ko, Y. D., Maniscalco, B., Carlson, C., Davidson, M., Doyle, W., Kuzniecky, R., Devinsky, O. and Halgren, E. (2017) 'Perceptual confidence neglects decision-incongruent evidence in the brain', *Nature Human Behaviour,* 1(7), s41562-017-0139.

Pfurtscheller, G. and Aranibar, A. (1979) 'Evaluation of event-related desynchronization (ERD) preceding and following voluntary self-paced movement', *Electroencephalogr Clin Neurophysiol,* 46(2), 138-46.

Pfurtscheller, G. and Berghold, A. (1989) 'Patterns of cortical activation during planning of voluntary movement', *Electroencephalogr Clin Neurophysiol,* 72(3), 250-8.

Pfurtscheller, G. and Lopes da Silva, F. H. (1999) 'Event-related EEG/MEG synchronization and desynchronization: basic principles', *Clin Neurophysiol,* 110(11), 1842-57.

Philiastides, M. G., Biele, G. and Heekeren, H. R. (2010) 'A mechanistic account of value computation in the human brain', *Proc Natl Acad Sci U S A,* 107(20), 9430-5.

Philiastides, M. G., Heekeren, H. R. and Sajda, P. (2014) 'Human scalp potentials reflect a mixture of decision-related signals during perceptual choices', *J Neurosci,* 34(50), 16877-89.

Philiastides, M. G., Ratcliff, R. and Sajda, P. (2006) 'Neural representation of task difficulty and decision making during perceptual categorization: a timing diagram', *J Neurosci,* 26(35), 8965-75.

Philiastides, M. G. and Sajda, P. (2006) 'Temporal characterization of the neural correlates of perceptual decision making in the human brain', *Cereb Cortex,* 16(4), 509-18.

Philiastides, M. G. and Sajda, P. (2007) 'EEG-informed fMRI reveals spatiotemporal characteristics of perceptual decision making', *J Neurosci,* 27(48), 13082-91.

Pisauro, M. A., Fouragnan, E., Retzler, C. and Philiastides, M. G. (2017) 'Neural correlates of evidence accumulation during value-based decisions revealed via simultaneous EEG-fMRI', *Nat Commun,* 8, 15808.

Pleskac, T. J. and Busemeyer, J. R. (2010) 'Two-stage dynamic signal detection: a theory of choice, decision time, and confidence', *Psychol Rev,* 117(3), 864-901.

Ploran, E. J., Nelson, S. M., Velanova, K., Donaldson, D. I., Petersen, S. E. and Wheeler, M. E. (2007) 'Evidence accumulation and the moment of recognition: dissociating perceptual recognition processes using fMRI', *J Neurosci,* 27(44), 11912-24.

Polania, R., Krajbich, I., Grueschow, M. and Ruff, C. C. (2014) 'Neural oscillations and synchronization differentially support evidence accumulation in perceptual and value-based decision making', *Neuron,* 82(3), 709-20.

Pouget, A., Drugowitsch, J. and Kepecs, A. (2016) 'Confidence and certainty: distinct probabilistic quantities for different goals', *Nat Neurosci,* 19(3), 366-74.

Preuschoff, K., Bossaerts, P. and Quartz, S. R. (2006) 'Neural differentiation of expected reward and risk in human subcortical structures', *Neuron,* 51(3), 381-90.

Rangel, A., Camerer, C. and Montague, P. R. (2008) 'A framework for studying the neurobiology of value-based decision making', *Nat Rev Neurosci,* 9(7), 545-56.

Rangel, A. and Hare, T. (2010) 'Neural computations associated with goal-directed choice', *Curr Opin Neurobiol,* 20(2), 262-70.

Ratcliff, R. (1978) 'A theory of memory retrieval', *Psychol Rev,* 85(2), 59.

Ratcliff, R., Philiastides, M. G. and Sajda, P. (2009) 'Quality of evidence for perceptual decision making is indexed by trial-to-trial variability of the EEG', *Proc Natl Acad Sci U S A,* 106(16), 6539-44.

Resulaj, A., Kiani, R., Wolpert, D. M. and Shadlen, M. N. (2009) 'Changes of mind in decision-making', *Nature,* 461(7261), 263-6.

Roitman, J. D. and Shadlen, M. N. (2002) 'Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task', *J Neurosci,* 22(21), 9475-89.

Rolls, E. T., Grabenhorst, F. and Deco, G. (2010a) 'Choice, difficulty, and confidence in the brain', *Neuroimage,* 53(2), 694-706.

Rolls, E. T., Grabenhorst, F. and Deco, G. (2010b) 'Decision-making, errors, and confidence in the brain', *J Neurophysiol,* 104(5), 2359-74.

Rolls, E. T., McCabe, C. and Redoute, J. (2008) 'Expected value, reward outcome, and temporal difference error representations in a probabilistic decision task', *Cereb Cortex,* 18(3), 652-63.

Rounis, E., Maniscalco, B., Rothwell, J. C., Passingham, R. E. and Lau, H. (2010) 'Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness', *Cogn Neurosci,* 1(3), 165-75.

Rushworth, M. F. and Behrens, T. E. (2008) 'Choice, uncertainty and value in prefrontal and cingulate cortex', *Nat Neurosci,* 11(4), 389-97.

Rushworth, M. F., Buckley, M. J., Behrens, T. E., Walton, M. E. and Bannerman, D. M. (2007) 'Functional organization of the medial frontal cortex', *Curr Opin Neurobiol,* 17(2), 220-7.

Sajda, P., Gerson, A. D., Philiastides, M. G. and Parra, L. C. (2007) 'Single-trial analysis of EEG during rapid visual discrimination: Enabling cortically-coupled computer vision' in Dornhege, G., Mullan, J. R., Hinterberger, T., McFarland, D. J. and Muller, K. R., eds., *Toward Brain-Computer Interfacing*, 1st ed., Cambridge, MA: MIT Press, 423-439.

Sajda, P., Philiastides, M. G. and Parra, L. C. (2009) 'Single-trial analysis of neuroimaging data: inferring neural networks underlying perceptual decision-making in the human brain', *IEEE Rev Biomed Eng,* 2, 97-109.

Samaha, J., Iemi, L. and Postle, B. R. (2017) 'Prestimulus alpha-band power biases visual discrimination confidence, but not accuracy', *Conscious Cogn*.

Selen, L. P., Shadlen, M. N. and Wolpert, D. M. (2012) 'Deliberation in the motor system: reflex gains track evolving evidence leading to a decision', *J Neurosci,* 32(7), 2276-86.

Shadlen, M. N., Britten, K. H., Newsome, W. T. and Movshon, J. A. (1996) 'A computational analysis of the relationship between neuronal and behavioral responses to visual motion', *J Neurosci,* 16(4), 1486-510.

Shadlen, M. N. and Kiani, R. (2013) 'Decision making as a window on cognition', *Neuron,* 80(3), 791-806.

Shadlen, M. N. and Newsome, W. T. (2001) 'Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey', *J Neurophysiol,* 86(4), 1916-36.

Siegel, M., Engel, A. K. and Donner, T. H. (2011) 'Cortical network dynamics of perceptual decision-making in the human brain', *Front Hum Neurosci,* 5, 21.

Smith, P. L. and Ratcliff, R. (2004) 'Psychology and neurobiology of simple decisions', *Trends Neurosci,* 27(3), 161-8.

Song, J. H. and Nakayama, K. (2009) 'Hidden cognitive states revealed in choice reaching tasks', *Trends Cogn Sci,* 13(8), 360-6.

Tallon-Baudry, C., Bertrand, O., Delpuech, C. and Pernier, J. (1996) 'Stimulus specificity of phase-locked and non-phase-locked 40 Hz visual responses in human', *J Neurosci,* 16(13), 4240-9.

Thut, G., Nietzel, A., Brandt, S. A. and Pascual-Leone, A. (2006) 'Alpha-band electroencephalographic activity over occipital cortex indexes visuospatial attention bias and predicts visual target detection', *J Neurosci,* 26(37), 9494-502.

Tosoni, A., Galati, G., Romani, G. L. and Corbetta, M. (2008) 'Sensory-motor mechanisms in human parietal cortex underlie arbitrary visual decisions', *Nat Neurosci,* 11(12), 1446-53.

Troje, N. F. and Bulthoff, H. H. (1996) 'Face recognition under varying poses: the role of texture and shape', *Vision Res,* 36(12), 1761-71.

Twomey, D. M., Murphy, P. R., Kelly, S. P. and O'Connell, R. G. (2015) 'The classic P300 encodes a build-to-threshold decision variable', *Eur J Neurosci,* 42(1), 1636-43.

Tzagarakis, C., Ince, N. F., Leuthold, A. C. and Pellizzer, G. (2010) 'Beta-band activity during motor planning reflects response uncertainty', *J Neurosci,* 30(34), 11270-7.

Tzagarakis, C., West, S. and Pellizzer, G. (2015) 'Brain oscillatory activity during motor preparation: effect of directional uncertainty on beta, but not alpha, frequency band', *Front Neurosci,* 9, 246.

Usher, M. and McClelland, J. L. (2001) 'The time course of perceptual choice: the leaky, competing accumulator model', *Psychol Rev,* 108(3), 550-92.

van den Berg, R., Anandalingam, K., Zylberberg, A., Kiani, R., Shadlen, M. N. and Wolpert, D. M. (2016a) 'A common mechanism underlies changes of mind about decisions and confidence', *Elife,* 5, e12192.

van den Berg, R., Zylberberg, A., Kiani, R., Shadlen, M. N. and Wolpert, D. M. (2016b) 'Confidence Is the Bridge between Multi-stage Decisions', *Curr Biol,* 26(23), 3157-3168.

van Dijk, H., Schoffelen, J. M., Oostenveld, R. and Jensen, O. (2008) 'Prestimulus oscillatory activity in the alpha band predicts visual discrimination ability', *J Neurosci,* 28(8), 1816-23.

van Maanen, L., Brown, S. D., Eichele, T., Wagenmakers, E. J., Ho, T., Serences, J. and Forstmann, B. U. (2011) 'Neural correlates of trial-to-trial fluctuations in response caution', *J Neurosci,* 31(48), 17488-95.

Van Vugt, M. K., Simen, P., Nystrom, L. E., Holmes, P. and Cohen, J. D. (2012) 'EEG oscillations reveal neural correlates of evidence accumulation', *Front Neurosci,* 6, 106.

Vickers, D. (1979) *Decision Processes in Visual Perception,* Academic Press.

Vickers, D. and Packer, J. (1982) 'Effects of alternating set for speed or accuracy on response time, accuracy and confidence in a unidimensional discrimination task', *Acta Psychol (Amst),* 50(2), 179-97.

Vickers, D., Smith, P., Burt, J. and Brown, M. (1985) 'Experimental Paradigms Emphasizing State or Process Limitations .2. Effects on Confidence', *Acta Psychol (Amst),* 59(2), 163-193.

Wallis, J. D. (2007) 'Orbitofrontal cortex and its contribution to decision-making', *Annu Rev Neurosci,* 30, 31-56.

White, T. P., Engen, N. H., Sorensen, S., Overgaard, M. and Shergill, S. S. (2014) 'Uncertainty and confidence from the triple-network perspective: voxel-based meta-analyses', *Brain Cogn,* 85, 191-200.

Wokke, M. E., Cleeremans, A. and Ridderinkhof, K. R. (2017) 'Sure I'm Sure: Prefrontal Oscillations Support Metacognitive Monitoring of Decision Making', *J Neurosci,* 37(4), 781-789.

Wyart, V., de Gardelle, V., Scholl, J. and Summerfield, C. (2012) 'Rhythmic fluctuations in evidence accumulation during decision making in the human brain', *Neuron,* 76(4), 847-58.

Wyart, V. and Tallon-Baudry, C. (2009) 'How ongoing fluctuations in human visual cortex predict perceptual awareness: baseline shift versus decision bias', *J Neurosci,* 29(27), 8715-25.

Yeung, N. and Summerfield, C. (2012) 'Metacognition in human decision-making: confidence and error monitoring', *Philos Trans R Soc Lond B Biol Sci,* 367(1594), 1310-21.

Yu, S., Pleskac, T. J. and Zeigenfuse, M. D. (2015) 'Dynamics of postdecisional processing of confidence', *J Exp Psychol Gen,* 144(2), 489-510.

Zemel, R. S., Dayan, P. and Pouget, A. (1998) 'Probabilistic interpretation of population codes', *Neural Comput,* 10(2), 403-30.

Zizlsperger, L., Sauvigny, T., Handel, B. and Haarmeier, T. (2014) 'Cortical representations of confidence in a visual perceptual decision', *Nat Commun,* 5, 3940.

Zylberberg, A., Barttfeld, P. and Sigman, M. (2012) 'The construction of confidence in a perceptual decision', *Front Integr Neurosci,* 6, 79.

Zylberberg, A., Fetsch, C. R. and Shadlen, M. N. (2016) 'The influence of evidence volatility on choice, reaction time and confidence in a perceptual decision', *Elife,* 5.

Zylberberg, A., Roelfsema, P. R. and Sigman, M. (2014) 'Variance misperception explains illusions of confidence in simple perceptual decisions', *Conscious Cogn,* 27, 246-53.