Confidence in Adaptive Decision-Making



Maja Friedemann

St John's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity Term 2023

In memory of my dad,

Hans-Heiner Friedemann,

who would have enjoyed to see me do this DPhil

more than anyone else.

Acknowledgements

First off, I would like to thank my supervisors, Nick Yeung and Dan Bang. Nick, for your steady support and guidance, your unique ability to gently nudge me back on track whenever I started overthinking, and for constantly encouraging me to recalibrate my own confidence. Dan, for being so generous with your time and advice, for always being there to toss around research ideas, and for your kind and motivating words that never failed to leave me more optimistic.

Thank you also to Claire Gillan and the Gillan gals for making my research visit to Trinity College Dublin an unforgettable experience. You all made me feel right at home and working with you has made my DPhil journey all the more exciting.

My work has been shaped by conversations with many brilliant and kind members of the Experimental Psychology department. To all past and present members of the ACC lab: Thank you for your companionship, insightful discussions, and support. Thank you to Matt, for being my rubber duck when I had silly coding questions and for patiently helping me with more complicated ones. Linda, for all your advice and for sharing your JavaScript code without which my experiments would have looked only half as pretty. Jasmine, for cheering me up throughout EEG decoding struggles. Samuel, for your humour and for always being around (especially on the weekends). Sammi, for helping me so kindly with my EEG work and sharing my determination to avoid MATLAB at all costs. Cody, for coffee breaks and dog walks

filled with philosophical discussions or monologues about yourself. And Jonas, for your aggressive enthusiasm and sharing the highs and lows of the DPhil.

I am also very grateful for the support I have received from individuals outside the EP department. Thank you to Alan Jagolinzer and Kai Ruggeri for reigniting my love for research. Josy, for telling me to '*just submit anything*' when I almost decided not to apply to Oxford after all and for long working sessions over Skype that kept me going during the pandemic and that continued far beyond lockdown times. Celine, for being one of the kindest people I know and for your feedback on my work. Elise, for imagined and real hiking expeditions and for being the best captain in all of SJC Boat Club's history. Thank you also to my family and my friends in Germany, in particular to my mum, Jan, Maggy, Anna, Sophie, Clara, Anaïs, and Jojo for always being there for me. And to Tariq, for knowing all my detrigger-buttons and simply making everything better.

This work was made possible by a Grand Union Doctoral Training Partnership Studentship, jointly funded by the Economic and Social Research Council and St. John's College Oxford. I am also grateful for the additional support from the COVID-19 Scholarship Extension Fund offered by the University of Oxford.

Abstract

The capacity for introspection and self-evaluation of our own thoughts and behaviours is a fundamental aspect of human cognition. Central to this faculty is the judgement of confidence in our beliefs and decisions, particularly in situations in which external feedback is sparse or absent. This thesis explores the critical role of confidence judgements in adaptive decision-making, substantiated by empirical research within the rigorously controlled environment of perceptual decision tasks. The first line of research probes the flexibility of confidence-guided information seeking behaviour. Findings illustrate a flexible interplay between confidence and information-seeking choices, contingent upon reward structures and the usefulness of information. However, the findings also unveil the boundaries of such flexibility, revealing a reliance on heuristic strategies, wherein a decrease in confidence generally prompts the need for more information, when dealing with more nuanced aspects of the choice environment. The second line of research investigates the context-sensitive role of confidence within social settings, uncovering distinct impacts of informational and normative social influences on confidence. The results indicate that in contexts devoid of feedback, individuals utilise others' confidence judgements as a benchmark to calibrate their own. Furthermore, under normative social influences, individuals strategically adapt their overt communication of confidence to suit the contextual demands. The third line of research examines the link between confidence biases and maladaptive behaviours, focusing on the example

of problem gamblers. This inquiry reveals that problem gamblers display heightened decision confidence, yet concurrently exhibit reduced self-esteem compared to a control group. Such overconfidence may be a contributing factor to perpetuating gambling behaviours, which highlights the necessity of well-calibrated confidence for adaptive behaviour. Collectively, these research lines emphasise that confidence is a multifaceted, context-sensitive mechanism, playing a key role in guiding adaptive behaviour. This thesis highlights the pervasive influence of confidence across various facets of human life — from everyday decision-making to psychopathology — thereby underlining its vital role in adaptive decision-making.

Contents

1 Introduction **1**

1.1 Confidence: an example of metacognition 2

1.2 Measuring metacognition 4

1.3 Neural substrates of metacognition 8

1.4 Metacognition guides adaptive behaviour 18

1.5 Metacognition in social settings 21

1.6 Metacognition in psychiatric disorders 28

1.7 Thesis outline 34

2 Confidence-Guided Information Seeking **37**

2.1 Introduction 39

2.2 Experiment 1 45

2.2.1 Methods 46

2.2.2 Results 49

2.2.3 Discussion 58

2.3 Experiment 2 62

2.3.1 Methods 64

2.3.2 Results 65

2.3.3 Discussion 71

2.4 General Discussion 72

3 Confidence Matching in Social Settings **79**

3.1 Introduction . . . 81

3.2 Methods . . . 90

3.3 Results . . . 97

3.4 Discussion . . . 110

4 Neural Correlates of Confidence Matching **117**

4.1 Introduction . . . 119

4.2 Methods . . . 124

4.3 Results . . . 130

4.4 Discussion . . . 150

5 Confidence Biases in Problem Gamblers **159**

5.1 Introduction . . . 161

5.2 Methods . . . 166

5.3 Results . . . 172

5.4 Discussion . . . 177

6 General Discussion **187**

6.1 Metacognition guides adaptive behaviour . . . 188

6.2 Metacognition in social settings . . . 191

6.3 Metacognition in psychiatric disorders . . . 193

6.4 Refining the model of metacognition . . . 195

6.5 Neural representation of confidence . . . 202

6.6 How we ask matters . . . 204

6.7 Conclusion . . . 205

Bibliography **207**

1

Introduction

Chapter Abstract

This opening chapter of my thesis establishes the foundation for the following empirical chapters which explore the role of confidence, a metacognitive judgement, in facilitating adaptive decision-making. In this chapter, I first give an overview of metacognition from its conceptualisation to measurement techniques and findings from work on metacognition within the cognitive neurosciences. This discussion serves to introduce key concepts and research themes common in metacognition studies. I then integrate the core themes of the thesis, considering the critical function of confidence in guiding adaptive behaviours, its role in social settings, and metacognitive abnormalities that have been linked to psychiatric disorders. The chapter concludes with an overview of the scope and outline of the thesis.

1.1 Confidence: an example of metacognition

Metacognition, as first conceived by Flavell (1979), is '*knowledge and cognition about cognitive phenomena*' (*p.* 906). That is, put simply, metacognition is thinking about thinking. For instance, the decision to set yourself a reminder requires metacognitive processing in that the decision is prompted by knowledge about the limitations of your own memory, and depends on how confident you are that you will remember a particular piece of information. Similarly, choosing which lecture content to revise prior to an exam employs metacognition as it depends on evaluating how confident you are in your own level of understanding of the lecture material. Metacognition plays an important role in adaptive behaviour and executive function across a range of cognitive domains. Understanding metacognitive processes is also relevant to many applied areas, such as diagnostic decisions in medicine or jury decisions in the courtroom, where subjective confidence estimates can have far-reaching consequences.

Metacognition initially received most attention in applied fields such as educational psychology (e.g., Baker, 1984; Nelson and Narens, 1990; Zimmerman, 1986) and organisational psychology (e.g., Bonaccio and Dalal, 2006; Sniezek and Van Swol, 2001) but has more recently been investigated through the lens of cognitive neuroscience (e.g., Fleming and Dolan, 2012; Vaccaro and Fleming, 2018). Within the realm of educational psychology, research is centered around students' metacognitive abilities to evaluate and modify their own learning strategies. For instance, studies have shown that even preschool children can regulate their own behaviour using basic metacognitive strategies and monitor their own learning activities, choosing tasks and objectives that best match their cognitive abilities (Bronson, 2000; Winne, 2017; Winne and Perry, 2000). In the context of organisational psychology, the emphasis of metacognition research is on its pivotal role in communication, teamwork,

and the efficacy of leadership within a corporate environment. For example, research in this area has explored the importance of confidence in leadership and group influence, suggesting an increased willingness to accept advice from a more confident source, regardless of its objective accuracy (Price and Stone, 2004; Sniezek and Van Swol, 2001). Cognitive neuroscience research targets the characterisation of computational and neural substrates underpinning metacognition. Not only can this research tell us something about the neurobiological basis of metacognition, but it may also help elucidate its underlying mechanisms and inform psychological theories of metacognition. This introductory chapter will initially focus on work in the cognitive neurosciences, as these studies illustrate many of the key theoretical challenges in the study of metacognition. As such, neuroscience research will act as a gateway to introduce essential concepts and research themes prevalent in the study of metacognition.

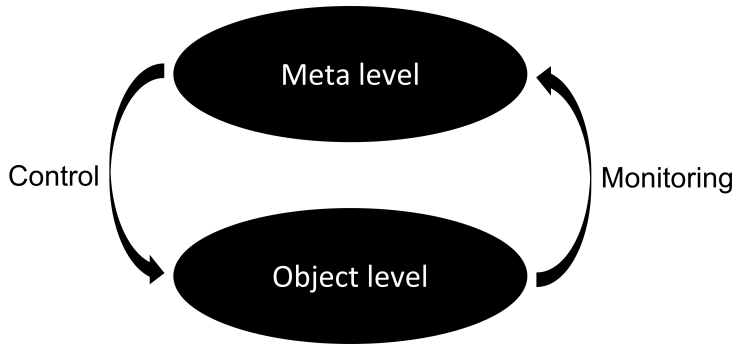


Figure 1.1: Model of metacognition adapted from Nelson and Narens (1990).

An influential model by Nelson and Narens (1990) conceptualises metacognition as a regulatory loop whereby object-level cognitive processes — such as memory, perception and decision-making — are monitored in a feedforward manner and then controlled in a feedback loop by meta-level cognitive processes encompassing higher-order representations of the lower object-level ones (see Figure 1.1). Two key aspects of metacognition research emerge from this model: understanding how and how well

people make metacognitive evaluations (monitoring), and understanding how they use these evaluations both individually and in groups (control). As this hierarchical layering of cognition may implicate a certain degree of self-awareness and self-reflection, metacognition has also been associated with consciousness (Koriat et al., 2006). Although some metacognitive processes may be explicit as in the examples above, they can also operate implicitly without conscious awareness (Shea and Frith, 2016). For instance, following error commitment, people often become more careful as shown by longer reaction times. Post-error adaptations may however occur outside awareness (Logan and Crump, 2010). Such expressions of metacognitive processes, not requiring verbal reports, also allow for studying metacognition in non-human animals (Terrace and Son, 2009) and pre-verbal infants (Goupil and Kouider, 2019). Whether this implicit type of self-monitoring should be classified as metacognition is debated (Proust, 2010). In this thesis, I mainly focus on explicit metacognitive processes.

1.2 Measuring metacognition

Research on metacognition typically requires participants to reflect on ongoing cognition or behaviour. In the laboratory, participants are usually asked to provide second-order confidence judgements about memories, decisions, or perceptions of the world. These second-order reports can be elicited prospectively or retrospectively. Classic examples of prospective metacognitive reports are Feelings of Knowing (Hart, 1965) and Judgements of Learning (Arbuckle and Cuddy, 1969). Feeling of Knowing ratings reflect a participant's belief about whether they will later be able to recognise the correct answer to a question or a problem from a list of alternatives. Judgement of Learning ratings reflect a participant's evaluation of whether they will be able to recall learned material in the future. In contrast, retrospective reports are elicited by probing participants' confidence in an answer they have already given,

or via post-decision wagering, whereby participants are asked how much they would bet that a decision they have made is correct (Fleming and Lau, 2014).

An individual's metacognition is typically quantified using measures such as metacognitive sensitivity, metacognitive bias, and metacognitive efficiency (Fleming and Lau, 2014). These measures probe the association between subjective metacognitive judgements and objective task performance. Metacognitive sensitivity, also called resolution, quantifies the ability to discriminate correct and incorrect judgements. Metacognitive bias, also called calibration, quantifies the overall level of confidence independent of metacognitive sensitivity (see Figure 1.2). Finally, metacognitive efficiency quantifies the level of metacognitive sensitivity taking into account first-order task performance, as the same person may display greater metacognitive sensitivity in a simple task compared to a difficult one or vice versa.

Metacognitive sensitivity can be calculated as the correlation between subjective ratings and objective performance, but is more often calculated using non-parametric approaches such as the Receiver Operating Characteristic (ROC) curve or approaches based on Signal Detection Theory (SDT) such as meta-d' (Fleming and Lau, 2014; Maniscalco and Lau, 2012). A correlational estimation of metacognitive sensitivity is via the Goodman-Kruskal gamma coefficient G (Nelson, 1987). The gamma coefficient is a rank correlation with responses ranked according to reported confidence bins and response accuracy. The gamma coefficient compares the total number of response pairs with a concordant ranking ($confidence_a > confidence_b$ and $accuracy_a > accuracy_b$) to those with a discordant ranking ($confidence_a > confidence_b$ but $accuracy_a < accuracy_b$).

ROC curves plot for each confidence value the probability of a confidence report being above that value for correct responses against the probability of a confidence report being above that value for incorrect responses. The area under the ROC curve

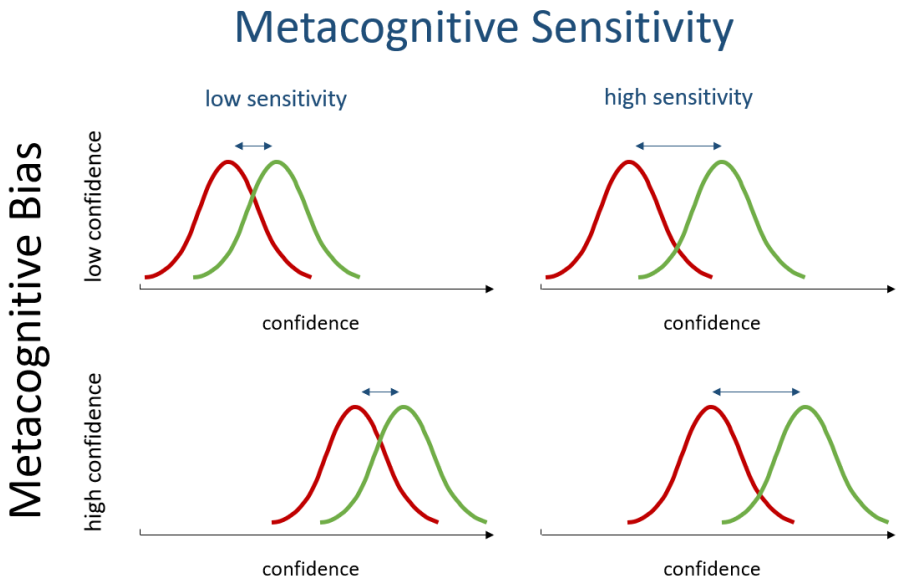


Figure 1.2: Each plot represents the distribution of confidence ratings for correct (green) and incorrect (red) choices. Metacognitive bias refers to the overall level of confidence expressed: the distributions on the top are centered around an overall lower reported confidence level, whereas the distributions on the bottom are centered around an overall higher reported confidence level. Metacognitive sensitivity refers to the separation between confidence distributions for correct and incorrect choices: the distributions on the left graphs are closer together and do not separate well between correct and incorrect choices, whereas the distributions on the right clearly distinguish between correct and incorrect choices.

(AUROC) is a measure of how well confidence ratings distinguish between correct and incorrect responses, whereby higher AUROC means higher metacognitive sensitivity, with a value of 0.5 indicating performance at chance level and an AUROC of 1 indicating perfect discrimination (see Figure 1.3). Meta-d' is a related measure that makes stronger assumptions about the underlying process generating metacognitive reports: it evaluates the type 1 d' (the sensitivity for discriminating stimulus alternatives, i.e. first-order object-level processes) that would have created the observed type 2 data if the participant had perfect metacognitive sensitivity, given the presumed relationship between these measures (Galvin et al., 2003). Metacognitive efficiency is typically assessed in terms of the ratio between meta-d' and type 1 d', called M-ratio (Fleming and Lau, 2014). An M-ratio equal to 1, signifies that all the first-order information is fully utilised in the second-order confidence judgement. An M-ratio smaller than 1, indicates that the metacognitive judgement does not utilise all the available information from the first-order response.

The optimal approach for measuring metacognitive performance remains a topic of active debate. Guggenmos (2021) recently scrutinized data from the Confidence Database (Rahnev et al., 2020), a collective repository of confidence-related datasets, and showed that no existing measure of metacognitive performance is entirely free from the influence of first-order performance (i.e., the general confound that people's metacognitive sensitivity degrades as their task performance drops, for example as task difficulty increases). Additionally, the reliability of these measures heavily relies on the combination of first-order performance and the number of trials in a task. Guggenmos (2021) further highlighted that the number of trials typically used in metacognition research does not meet the standard for acceptable test-retest reliability. A subsequent evaluation of current metacognitive measurement methods conducted by Rahnev (2023) also emphasised that no current measure is perfect and that depending on the specifics of the experiment, certain measures might be more

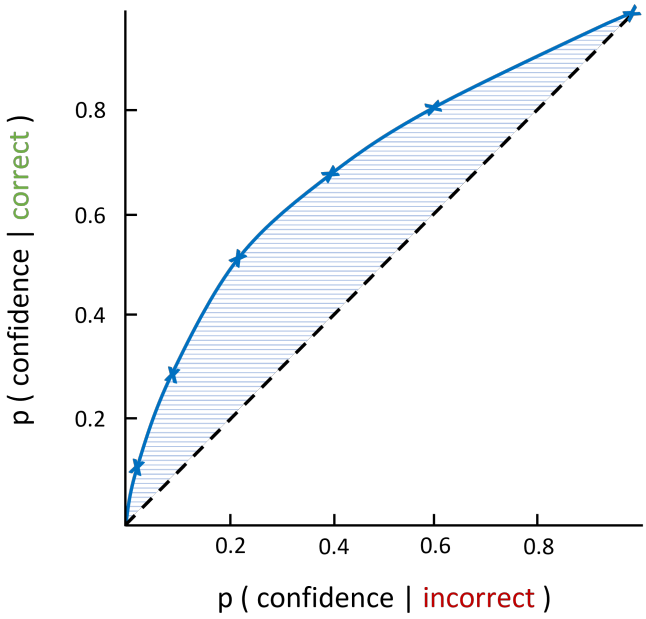


Figure 1.3: Example Receiver Operating Characteristic (ROC) curve. The curve plots the hit rate against the false alarm rate at various confidence thresholds. The shaded area under the curve (AUROC) provides a measure of metacognitive sensitivity.

fitting than others. Future approaches to measuring metacognitive ability may lie with model-based measures. For example, recent model-based measures have been proposed by Mamassian and de Gardelle (2021) and by Desender et al. (2022). In the ensuing chapters, presenting my empirical work, I focus on mean confidence and the variation of confidence reports across trials. Although I provide AUROC measures to give a general sense of metacognitive performance in the tasks, I do not analyse metacognitive sensitivity measures in my experiments.

1.3 Neural substrates of metacognition

Identifying the neural substrates of a metacognitive process comes with multiple challenges. One challenge is to separate metacognitive from object-level processes in neural data as these processes often involve the same task variables (Bang and Fleming, 2018). For instance, if we find neural activity that varies with a person's

reported confidence in a decision, how can we be certain that this activity reflects the coding of confidence itself rather than variation in the strength of evidence supporting a decision at the object-level? Indeed, based on findings from experiments conducted with rats and monkeys, it has been put forward that confidence may be computed automatically as part of the probabilistic computations of the decision process (Kepecs et al., 2008; Kiani and Shadlen, 2009), suggesting it may be naive to seek to separate cognitive and metacognitive representations. Yet more complexity is added by the fact that metacognitive judgements depend on multiple cues, such as Feelings of Knowing in memory reflecting both the familiarity of the question and the accessibility of possible answers (Koriat and Levy-Sadot, 2001; Schwartz and Díaz, 2014). If so, should we expect to see distinct neural correlates of each of these cues, an integrated representation of all cues combined, or both types of representation? Responding to these challenges, researchers have employed carefully-designed experiments to dissociate metacognitive from object-level processes across various cognitive domains. I next review key themes and debates emerging from this research as an introduction to core questions that have been considered in metacognition research more widely.

Metacognitive processes can be dissociated from object-level ones. Early evidence for specific brain regions being involved in metacognitive (and not object-level) processes comes from studies of patients who suffered brain damage to the prefrontal cortex (PFC) — a plausible neural substrate for metacognition given its longstanding association with higher cognitive functions (Friedman and Robbins, 2021). Shimamura and Squire (1986) and Janowsky et al. (1989) showed that, while amnesic patients displayed object-level memory deficits when compared to healthy controls, only patients with frontal lobe dysfunction appeared to have selective deficits in their meta-memory as manifested in impaired Feeling of Knowing judgements. These patients made poor predictions about their ability to recognise

the correct answer to general knowledge questions in cases where they could not currently recall that answer. The unique impairments in meta-memory performance compared to other amnesic patients suggest that the deficits do not result from general memory impairments. Kao et al. (2005) presented further evidence that distinguishes brain regions concerned with actually retrieving a memory from regions monitoring confidence in those memories: they found activity in the ventromedial PFC to vary with participants' predictions of whether they would remember versus forget presented scene images, independent of their actual first-order memorisation success, whereas activity in the medial temporal lobe showed the opposite pattern. Fleming et al. (2010) provided converging evidence using a different imaging technique and in a different task domain: correlating structural magnetic resonance imaging (MRI) data with task performance, they found that gray-matter volume in a region in right anterior PFC predicted participants' metacognitive sensitivity in a perceptual decision task, after carefully controlling for individual task performance. Meanwhile, studies leveraging the temporal resolution of electroencephalography (EEG) have shown that people's reported confidence in their decisions depends critically on continued processing after the initial choice, and that this post-decision processing is reflected in neural activity distinct from (indeed, anti-correlated with) activity seen during the object-level decision (Boldt and Yeung, 2015; Murphy et al., 2015).

Temporal focus of metacognition. An interesting pattern in neuroscientific studies of metacognition is a distinction between brain areas involved in prospective and retrospective metacognitive monitoring (i.e., predictions of future performance versus evaluations of past performance). For example, Schnyer et al. (2004) found that patients with damage in the right ventromedial PFC made poor prospective predictions about their ability to recognise the correct word to complete a previously-learned sentence. However, after attempting to recall sentence-completing words,

they were as accurate as controls in retrospectively judging whether the words recalled were the correct ones. Conversely, Pannu et al. (2005) found deficits in retrospective confidence judgements but unaffected performance in prospective Feeling of Knowing judgements in patients with damage in the lateral PFC performing a face memory task. Taken together, these findings suggest a separation in PFC between medial regions that are involved in predicting how well one will perform in the future and lateral regions that are involved in monitoring how well one has performed in the past (Fleming and Dolan, 2012). Subsequent neuroimaging studies have provided evidence consistent with this neuropsychological double dissociation (Kao et al., 2005; Lapate et al., 2020).

Domain-generality of metacognition. Similar to this separation of metacognitive judgements based on their temporal focus (prospective versus retrospective), there is a debate about whether separate neural substrates support metacognitive judgements across cognitive domains (e.g., memory versus perception; Vaccaro and Fleming, 2018). A lesion study assessed patients' metacognition in perception and memory to test whether metacognition is globally supported by the anterior PFC across different cognitive object-level domains (Fleming et al., 2014). Whereas patients with lesions to the anterior PFC showed impairments in perceptual metacognitive accuracy, their meta-memory accuracy remained intact, suggesting that there may be domain-specific neural substrates of metacognition. This notion is further supported by neuroimaging studies in healthy participants with evidence from resting-state connectivity analyses (Baird et al., 2013), measurements of variation in white matter microstructure (Baird et al., 2015), and MRI using voxel-based morphometry analyses (McCurdy et al., 2013) and cortical thickness mapping (Valk et al., 2016). Reconciling these findings with studies arguing for domain-general metacognitive processes (e.g., De Gardelle and Mamassian, 2014; Fleck et al., 2006; Schraw, 1996), Morales et al. (2018) proposed the coexistence of domain-specific

and domain-general aspects of metacognition. This idea is also supported by computational accounts of metacognition which posit the distinct functional roles of domain-specific representations of certainty versus domain-general metacognitive representations of confidence (Fleming and Daw, 2017; Pouget et al., 2016).

Supramodality of metacognition. There is now converging evidence that at least some metacognitive processes operate across cognitive domains. For example, Faivre et al. (2018) found that metacognitive efficiency was correlated across tactile, auditory, visual, and audiovisual tasks. In line with previous work on EEG markers of error detection (Falkenstein et al., 1991), they also identified EEG markers of confidence shared across modalities — supporting the existence of supramodal (domain-general) metacognitive processes. De Gardelle et al. (2016) showed that participants can accurately compare confidence estimates across auditory and visual tasks — again supporting the existence of supramodal metacognitive processes. Such a '*common currency*' representation of confidence estimates has also been demonstrated in animal models. Masset et al. (2020) trained rats to perform olfactory and auditory perceptual decision tasks and used their willingness to wait for a performance-dependent reward as a proxy for confidence in their decisions. Employing single-cell recordings, they identified neurons in orbitofrontal cortex (OFC) that encoded decision confidence irrespective of the sensory modality.

Implicit and explicit metacognition. A further pattern that has emerged is a distinction between posterior/medial and anterior/lateral parts of the PFC. Activity in more posterior/medial regions such as dorsomedial PFC has been implicated in uncertainty monitoring across a wide range of tasks encompassing different cognitive domains (Yeung, 2014), whereas activity in anterior/lateral regions, such as the lateral frontal pole and BA10, seems to be linked to more nuanced subjective confidence reports (Bang et al., 2020; Bang and Fleming, 2018). For instance, Morales

et al. (2018) reported activity in posterior PFC as a function of generic uncertainty across memory and perception tasks, whereas activity in anterior PFC was sensitive to the domain in which metacognitive signals were probed. In a study by Fleming et al. (2018), participants were presented with additional evidence after having made a decision, which supported or contradicted the initial choice. Activity in posterior/medial PFC was related to this objective post-decision evidence, whereas activity in a more anterior/lateral region appeared to encode the impact additional evidence had on subjective confidence judgements.

In line with these findings, Qiu et al. (2018) suggested that there may be two stages of metacognitive processes: the first stage reads out generic uncertainty information from object-level processes and the second stage then transforms this information into a common sense of confidence. This notion is supported by a study conducted by Shekhar and Rahnev (2018) who observed complementary effects of transcranial magnetic stimulation on posterior and anterior PFC regions. Posterior PFC stimulation reduced participants' reported confidence overall — in line with these regions encoding the strength of sensory evidence. By contrast, anterior PFC stimulation affected metacognitive sensitivity — in line with these regions using information about signal strength (and potentially other sources of information) to generate explicit confidence judgements. Going back to the idea that some metacognitive processes can operate implicitly without conscious awareness, the lower-level processes in posterior parts of the PFC may operate without awareness, whereas those operating in the anterior PFC may be more aware allowing for the reporting and more flexible and nuanced use of conscious confidence judgements (Bang et al., 2020; Gherman and Philiastides, 2018).

Have we identified neural substrates of metacognitive monitoring? The neuropsychological and neuroimaging studies reviewed above reveal consistent in-

volvement of prefrontal regions in metacognitive processes, with rostrolateral, ventromedial, and dorsomedial regions each implicated in many studies (see Figure 1.4). However, no single region has been consistently identified and hypothesised functional sub-divisions are yet to be fully replicated. Moreover, regions outside of the core network shown in Figure 1.4 have also been implicated: a formal meta-analysis of 47 neuroimaging studies identified a domain-general metacognition network including the medial and lateral PFC, as reviewed above, but also the precuneus and the insula (Vaccaro and Fleming, 2018), and EEG correlates of confidence exhibit a parietal rather than frontal focus (Boldt and Yeung, 2015). Activation foci are not even limited to the cortex, with Hebart et al. (2016) and Gherman and Philiastides (2018) reporting that activity in ventral striatum also tracked participants' reported confidence. Similarly, although studies in animals have observed predictable correlates of confidence (e.g., as reflected in animals' willingness to wager or wait for reward) in activity in PFC (e.g., Kepecs et al., 2008; Middlebrooks and Sommer, 2012; Miyamoto et al., 2017), corresponding neural correlates have been identified in regions as widespread as the pulvinar nucleus (Komura et al., 2013) and midbrain dopamine system (Lak et al., 2017).

This heterogeneity raises important questions about our conceptual understanding of metacognition and, correspondingly, what expectations we should have about underlying neural substrates. Current theories typically characterise confidence as a general representation of an agent's probability of success that is abstracted from the specific object-level task being performed, permitting flexible and generalised metacognitive control of thought and behaviour (Yeung and Summerfield, 2012). As such, we should be able to identify a reproducible representation of a metacognitive signal to be observed in the same brain area across multiple studies. However, if we consider the complexity of metacognition, whereby metacognitive judgements may depend on multiple cues and be used flexibly depending on the situation, then per-

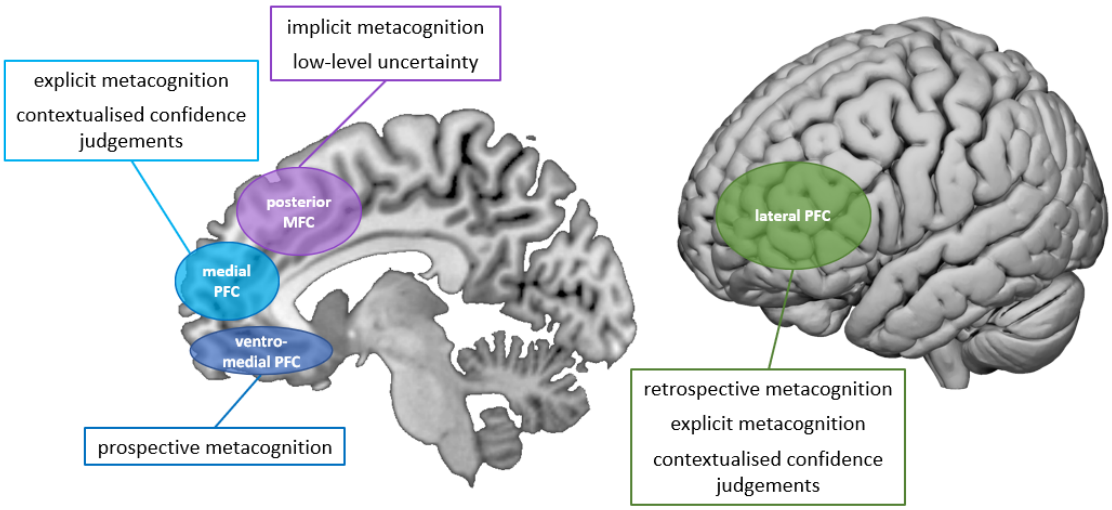


Figure 1.4: Medial (left) and lateral (right) view of the human brain highlighting hypothesised functional roles in metacognition. Whereas posterior medial frontal cortex — including dorsal anterior cingulate cortex and pre-supplementary motor area — seems to be involved in implicit metacognitive processes, medial and lateral prefrontal cortex — including the lateral frontal pole and brodmann area 10 — have been implicated in more explicit metacognitive representations.

haps finding activity in different areas may be just what we would expect (Schwartz and Díaz, 2014). In this case, determining the neural substrates of metacognition requires first disentangling the various processes that underlie any particular metacognitive judgement. In addition, underlying neural activity may vary according to perhaps subtle aspects of task demands and may also vary across individuals in a manner that challenges standard cognitive neuroscience approaches that look for statistical consistency across participants and studies.

What is the nature of metacognitive representations? As reviewed above, PFC seems to play a critical role in supporting metacognitive processes but, within this broad region of cortex, different studies report different patterns of neural activity in relation to metacognitive processes. In part, this heterogeneity might reflect the variety of criteria used to evaluate whether something constitutes a neural representation of metacognition. As a basic criterion, if metacognitive neural activity

reflects the monitoring of object-level processes, putative indices of metacognitive processes should correlate with individuals' probability of being correct in a given task. Beyond this, the activity should predict behaviours that are based on metacognitive monitoring, such as confidence reports, decision wagers, or error corrections. Lesions or causal interventions to brain areas processing metacognitive evaluations should interfere with such behavioural manifestations while leaving object-level processes intact. Another criterion one may adopt is that there should be a general abstract representation of confidence that operates between any kind of input (e.g., different cognitive domains or sensory modalities) and any kind of output (e.g., betting on choices or communicating confidence to a social partner; Figure 1.5A). However, even with this variety of possible definitions, most current theories of metacognition would predict some underlying shared representation of the confidence computation itself. The inconsistencies of current findings with this notion may be of methodological or theoretical nature.

On a methodological level, perhaps operational constraints mean that the common underlying representation of confidence is not apparent across various studies. Each individual study identifies neural activity as a set of contrasts that necessarily entail object-level differences relating to the stimuli presented, required responses, and indices of metacognition used. Setting up experiments in this way might limit our ability to identify the shared neural activity between contrasted conditions: across various studies, there may be shared neural activity, but the overlapping population of neurons encoding abstract confidence may not be detected as being most active or as being best at decoding inputs to or outputs from a confidence computation. A truly general abstract confidence representation would require utmost compression of information allowing for flexible use depending on the circumstances such that in any given situation, it would likely not be the richest representation of all task-relevant variation.

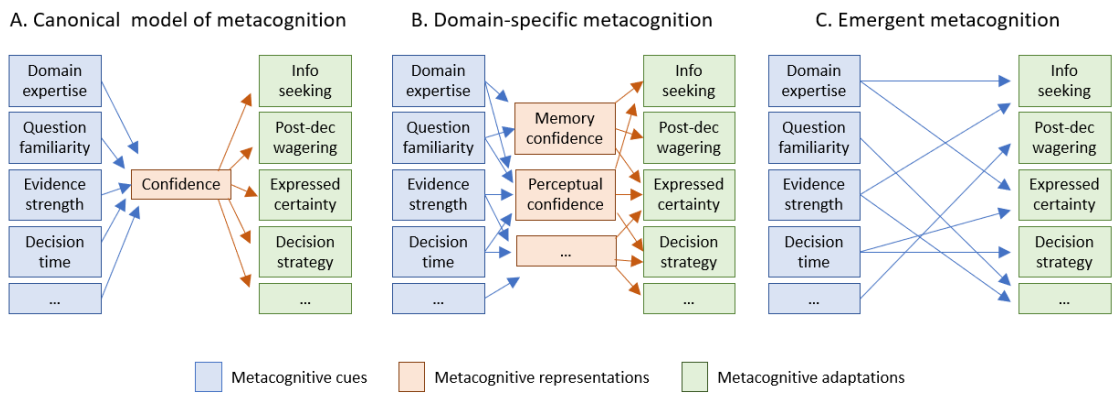


Figure 1.5: Models of metacognition. Each panel indicates a different theoretical relationship between cues that inform metacognitive processes, metacognitive representations (or their absence), and strategic control adjustments. In panels (B) and (C), the lack of full connectivity between processing stages captures the intuition that cues may weight differently across different metacognitive judgements and different forms of metacognitive control.

On a theoretical level, perhaps not every metacognitive experience derives from the same shared underlying representation. For instance, different representations may exist for different cognitive domains and only share some resemblance leading to a common subjective experience (see Figure 1.5B). That is, confidence might be represented in a different way neurally than our psychological theory may have us assume. Another possibility is that confidence may not in fact exist as commonly theorised, as an abstract representation that combines multiple cues into an integrated, explicit estimate of the reliability of object-level processes. Perhaps, instead, those multiple cues feed directly into a variety of strategic adaptations, without the need for a common representation or general mediation (see Figure 1.5C). It remains a key challenge for cognitive neuroscience research to unveil if an explicit confidence representation akin to currently dominant psychological theory does in fact exist in the brain, what this representation looks like, and where it is instantiated in the brain. A substantive gap persists between current theoretical accounts of metacognition – which emphasise explicit, generalised, abstract representations of confidence that guide flexible control of behaviour – and our current understanding of the neu-

ral underpinnings of this adaptive control. There remains substantial and currently unexplained variability in identified neural correlates both within core brain regions as well as the wider network of regions that might support metacognitive monitoring and control.

Having reviewed work on the neural basis of metacognitive monitoring and control processes, which illustrate key themes within metacognition research, the following three sections will present research that provides a direct foundation for the empirical studies reported in this thesis. These sections discuss work on metacognition guiding adaptive behaviours, metacognition in social settings, and metacognition in psychopathology, respectively.

1.4 Metacognition guides adaptive behaviour

The previous section reviewed key themes in metacognition research through the lens of investigations of the underlying neural mechanisms. In particular, we have seen that cognitive neuroscience has convincingly demonstrated that metacognitive processes can be distinguished from object-level operations in core cognitive domains. Neural substrates have been identified where activity correlates with, and lesions impair, metacognitive judgements but not object-level task performance, and anterior/lateral regions of PFC are consistently implicated in these processes. A fundamental question remains, however: why do we even have metacognition?

Emerging from the reviewed studies is the notion that metacognition enables flexible and adaptive control of behaviour. By being able to monitor one cognitive process (like our memory or decision-making), we can then apply this knowledge to exert control within a different cognitive process. For instance, recognising the limitations of our own memory may lead us to decide to write a grocery list in order to avoid forgetting any items. Indeed, metacognitive processes are important to initiate

self-regulatory actions, learn from mistakes, and optimise behaviour. In memory, Feelings of Knowing prompt sustained attempts at memory retrieval (Costermans et al., 1992; Koriat and Goldsmith, 1996), Judgements of Learning guide choices about which information should be refreshed and restudied (Nelson et al., 1994), and general evaluations of our own mental abilities and their shortcomings play an integral role in deciding when to set external reminders (Gilbert, 2015).

In decision-making, a critical task involves finding the right balance between gathering sufficient evidence to make an informed decision, while avoiding to spend unnecessary time and effort seeking information that does not further improve decision quality. Confidence, due to its reliable prediction of accuracy, has been proposed as a tool to directing the amount of evidence required prior to decision-making. Particularly, high confidence corresponds to a higher probability of making an accurate decision compared to situations where confidence is low, implying the need for more evidence collection before finalising a decision in the latter case (Yeung and Summerfield, 2012). This concept has been empirically supported through the finding that confidence predicts information seeking, such as reviewing evidence (Desender et al., 2018) and asking for advice (Pescetelli et al., 2021), even when controlling for objective performance accuracy, indicating the causal role of confidence judgements in deciding when to seek out information in an efficient manner.

Metacognitive processes not only improve immediate decision-making but also shape behavioural strategies for future decisions, for example by prompting the adoption of a more cautious approach following errors (Rabbitt and Rodgers, 1977) and low confidence decisions (Van den Berg et al., 2016). Van den Berg et al. (2016) examined the role of confidence in multi-stage decision-making, whereby tasks require a series of decisions to reach a goal. In their study, participants made two sequential perceptual decisions and were rewarded only if both were correct. The speed and

accuracy of these decisions were explained by the accumulation of noisy evidence up to a decision bound. Participants adjusted their speed-accuracy trade-off by increasing the decision bound for the second decision in proportion to their confidence in the first. Importantly, by relying on metacognitive signals, these behavioural adjustments do not necessitate any explicit external feedback (Guggenmos et al., 2016). On a longer timescale still, individuals can aggregate confidence judgements for multiple individual decisions over time into global beliefs about their overall ability and use these summary estimates to guide task selection (Carlebach and Yeung, 2020) and performance expectations for previously encountered (Boldt, Blundell, et al., 2019) or unacquainted tasks (Rouault, McWilliams, et al., 2018).

In value-based decision-making, a process whereby individuals make decisions based on the perceived worth or '*value*' of different choice options, people are more likely to stick to a decision if they previously experienced high confidence on the same choice (Folke et al., 2016). When deciding between options in complex environments, individuals must choose between exploiting the option they believe is best, or exploring new alternatives in hopes of a better outcome (Sutton and Barto, 2018). Boldt, Blundell, et al. (2019) showed that people use their subjective degree of certainty about the value of different options to decide between exploration and exploitation strategies. This was evident in the tendency to explore more when certainty about the value of different options was low, as this suggests more to learn about the environment.

Recent studies of the neural substrates of metacognitive control have convincingly linked behavioural adjustments to supposed metacognitive representation. In animal studies, reversible lesions to regions coding for confidence impair or abolish strategic behaviours such as adaptive waiting and post-decision wagering (Lak et al., 2014; Miyamoto et al., 2017). In human decision-making, EEG correlates of confidence

predict information seeking choices (Desender, Murphy, et al., 2019). In their fMRI study, Qiu et al. (2018) found overlapping correlates of confidence reports in very different tasks (perceptual decisions and Sudoku puzzles) in dorsomedial PFC and insula, but distinct activity patterns in anterior PFC that they linked to the different strategic adjustments available in the tasks. These studies highlight the importance of considering control processes in cognitive neuroscience studies, both as a central (and so far somewhat neglected) aspect of metacognition, as well as a potentially critical factor producing variability in activation foci across studies with different methodologies (Boldt and Gilbert, 2022).

Overall, metacognition plays a key role in guiding adaptive behaviour. Acting as an internal monitoring and control system, it allows us to self-regulate actions, learn from our mistakes, and optimise our performance in various cognitive domains. Confidence emerges not just a byproduct of cognition but a critical element that actively shapes our decisions and behaviours.

1.5 Metacognition in social settings

In addition to directing individual adaptive behaviour, metacognition serves a critical functional role in enabling us to articulate our experiences and justify our actions to others (Frith, 2012). In fact, Shea et al. (2014) argue that a key distinction between humans and other animals is our ability to not only implicitly reflect on our own cognitive processes and mental states but to also explicitly do so. This ability gives rise to the unique human capability to share representations of our own cognitive processes with others, thereby facilitating adaptive group behaviour and the building of a cumulative culture. Furthermore, Frith and Frith (2012) propose that metacognition enables us to recognise that our mental representations of the world are just that — representations. This insight leads us to understand that other

minds might possess different representations, and that our own could potentially be incorrect. Possessing these insights provides us with the unique opportunity to discuss and compare our reflections with others, ultimately contributing to the development of a shared perspective.

Studies of group decision and eyewitness testimony have explored how individuals' confidence reports are taken by others as a cue to their reliability (Bonaccio and Dalal, 2006; Sauer and Brewer, 2015). Throughout various life situations, even when the ultimate decision falls upon ourselves, we frequently gather insights from others before finalising our choice. For example, we may seek the advice of mentors, financial advisors, or friends and family members who have experience in the particular matter at hand. In these contexts, studies suggest that individuals typically adopt a confidence heuristic, attributing greater knowledge and competence to those who project high confidence levels, hence favouring their advice (Price and Stone, 2004; Pulford et al., 2018; Yates et al., 1996). The degree of confidence displayed by an advisor is often interpreted as a measure of their expertise. Consequently, advisors who exude more confidence tend to be trusted more and their advice is followed more frequently in comparison to less confident ones (Sniezek and Van Swol, 2001).

Recent research by Gaertig and Simmons (2023) adds a nuanced layer to the understanding of the impact of confidence on advice taking by investigating whether people are more or less likely to follow numerical advice that communicates uncertainty in the form of a confidence interval. Participants were more likely to choose the advisor's forecast over their own when the advice was accompanied by a confidence interval. These results were consistent across different measures of advice following and did not depend on the width of the confidence interval (75% or 95%), advice quality, or on whether people had information about the advisor's past performance. Thus, it seems that advisors who can communicate uncertainty effectively

might be more persuasive.

The impact of advisor confidence has also been researched within the legal domain. The weight placed upon eyewitness testimonies, a key element in suspect identification, is often heavily influenced by the confidence displayed by the eyewitness (Brewer and Hupfeld, 2004; Tenney et al., 2007). Witnesses who exude high confidence levels are generally viewed as more reliable such that their testimonies tend to have a stronger influence on jury decisions (Penrod and Cutler, 1995). Tenney et al. (2007) not only confirmed that high confidence testimonies have more influence than low confidence ones, but also showed that a witness expressing high confidence, if later proven incorrect, suffers a larger credibility loss than a less confident witness making the same error. This finding underscores the importance of confidence calibration.

A study by Kappes et al. (2020) however notes potential biases in interpreting the confidence levels displayed by others, providing evidence for a mechanism related to confirmation bias, whereby people tend to discount information that undermines past choices and judgements. The study found that whereas individuals adequately used the strength of others' confirmatory opinions to update their own confidence in judgements, they failed to do so for disconfirmatory opinions. This bias was related to a reduced neural sensitivity in the posterior medial prefrontal cortex to the strength of others' opinions when those opinions were different from one's own.

Beyond advice seeking, sharing metacognitive evaluations with others can facilitate adaptive group behaviour. When making joint decisions, individuals can determine who among themselves is most likely to know the correct answer by exchanging and comparing their confidence levels (Pescetelli et al., 2016). Zarnoth and Sniezek (1997) showed that an individual's self-reported confidence directly impacts their influence within a group, with more confident members having greater influence on the

collective decision of a freely interacting group. Bahrami et al. (2010) showed that dyadic decision-making performance can be well described by a weighted-confidence sharing model, whereby the individual responses are weighted by each agent's confidence in order to arrive at a joint decision. They discovered that when paired observers in a perceptual task could communicate the reliability of their choices (that is, their confidence judgements), their combined performance exceeded their individual accuracy levels. Linguistic analyses revealed that dyads who more quickly developed a shared set of descriptions to communicate their level of confidence reaped a greater group advantage (Fusaroli et al., 2012; but see Bang et al., 2017; Koriat, 2012 for potential issues with blind alignment of explicit confidence estimates). Sharing metacognitive representations may even compensate for missing objective feedback in group decision-making (Bahrami et al., 2012). The collective benefits of confidence sharing have been demonstrated not only in perceptual tasks but also in more complex cognitive tasks such as object-recognition (Brennan and Enns, 2015), enumeration (Bahrami et al., 2013; Brennan and Enns, 2015), and knowledge-based questions (Koriat, 2012).

Evidently, the sharing of confidence estimates serves as an effective strategy for cooperative decision-making. Although it is generally observed that a person's stated confidence level corresponds proportionately to the likelihood of their decision or belief being accurate, attributes of the confidence distribution, such as the average confidence, vary across individuals (Ais et al., 2016). Numerous factors such as mental health (Fu et al., 2012; Huq et al., 1988), physiological arousal (Allen et al., 2016; Jönsson et al., 2005), current mood (Koellinger and Treffers, 2015), gender (Lundeberg et al., 1994), profession (Broihanne et al., 2014), personality (Campbell et al., 2004; Pallier et al., 2002; Schaefer et al., 2004), and cultural background (Mann et al., 1998) have been shown to influence confidence reports. Consequently, an identical absolute expression of confidence might hold different implications under

various conditions and for different individuals. Hence, for the confidence heuristic — using confidence as a reliability indicator (Price and Stone, 2004) — to function as a successful communication strategy, it is crucial to mitigate confidence discrepancies that stem from circumstantial factors while preserving the variance induced by task-relevant signals (Bang et al., 2017). The calibration of confidence is also subject to individual differences (Fleming et al., 2010). Interactions involving individuals with poor confidence calibration can even result in a collective loss (Pescetelli et al., 2016). The benefits of joint decision-making have also been shown to be compromised by the adoption of an equality bias, whereby dyads tend to assign undue weight to the less accurate member and insufficient weight to the more accurate one, a tendency that persists even when feedback is provided on every trial (Mahmoodi et al., 2015).

Interestingly, research shows that when it comes to intellectual tasks, confidence does not always hold the highest sway in determining group decisions. Instead, compelling arguments can win over confidently expressed ones. A study by Trouche et al. (2014) demonstrated that the accuracy of information outweighs confidence when it comes to persuading others. The group adopted the correct answer in over 80% of the cases, even when the member proposing the correct solution was not the most confident one. Moreover, initial confidence did not predict susceptibility to accepting a contrary argument and changing one's mind. However, for good arguments to win over confidence, the correct conditions must be in place: the group members who cannot solve the problem themselves must have sufficient insight in order to recognise and accept the correct solution proposed by another member, and the correct member must be able to effectively communicate their solution to others.

The studies reviewed above have demonstrated how confidence plays a pivotal role in social contexts, as it is often leveraged to seek and interpret advice, as well as to facilitate group decision-making. Beyond this use of confidence in social situa-

tions, the social context has also been shown to shape an individual's self-directed metacognitive appraisals. For example, findings from a study by Mahmoodi et al. (2019) suggest that people use others' confidence judgements as a benchmark to calibrate their own. This was reflected in increased self-confidence when observing a less well performing partner, whereas conversely observing a more accurate partner resulted in diminished confidence. This finding indicates that our confidence in our own abilities and judgements is not solely based on our actual performance, but is also influenced by a comparative evaluation of others' performances. Similarly, Wittmann et al. (2016) found that people's self-assessments of ability were influenced by the performance of others with whom they were interacting. This influence varied depending on whether the context was cooperative or competitive. In a cooperative context, interacting with high performers enhanced individuals' own confidence, whereas the opposite effect was observed in a competitive context. The study further found that these influences not only affected subjective self-assessments but also impacted how people objectively adjusted their performance in subsequent tasks. This highlights that our confidence judgements can be swayed by the performance of others and the nature of our interaction with them, affecting both our confidence in our abilities and our actual performance.

A study by Will et al. (2017) focused on the neural and computational mechanisms underlying fluctuations in self-esteem as a result of social feedback. The research presented a computational model that correlates the difference between expected and received social feedback (termed social prediction errors) with activity in the ventral striatum/subgenual ACC. Changes in self-esteem resulting from these errors were associated with activity in the vmPFC. This study underscores the fact that our confidence judgements (in this case, self-esteem) are significantly influenced by the feedback we receive from others, showing the role of social influence in shaping our self-perceptions.

The influence of social dynamics on metacognition is further evidenced by a study on tip-of-the-tongue states (a powerful and tantalising sense that one knows something that cannot be recalled momentarily) conducted by Rousseau and Kashur (2021). They found that participants reported more tip-of-the-tongue states when recalling in small groups than when recalling alone. This effect did not increase correct recall, suggesting that the increase in tip-of-the-tongue states was not related to retrieval process improvement but possibly to the metacognitive appraisal of group recall efficiency. This suggests that the mere presence of others can heighten our perceived proximity to successful recall, thus influencing our confidence judgements. The study suggests that this could be due to an enhanced sense of recall imminence, brought on by the metacognitive assumption that group recall is more likely to be successful than individual recall.

In sum, metacognition holds a central role in navigating social exchanges and orchestrating group decision-making dynamics. Confidence projected by individuals typically influences trustworthiness perceptions, sharing confidence judgements can enhance the efficacy of joint decision-making, and our self-confidence evaluations can be shaped by the social context we find ourselves in, as well as the feedback we receive from others.

the literature reviewed above considers '*social*' within metacognition as encompassing diverse interactions, dynamics, and contexts, my empirical work reported in Chapters 3 and 4 of this thesis adopts a more narrow scope. In my experiments, '*social*' pertains to situations where another individual's input shapes a participant's metacognitive judgements. This definition hones in on two key facets of '*social*': the presence of another and the exchange of metacognitive information, such as confidence levels. Although my approach focuses on this basic understanding, I acknowledge the broader, intricate dimensions of '*social*' presented in previous research. My

intention is to establish a base for further exploration into the role of metacognition within more complex interpersonal dynamics.

1.6 Metacognition in psychiatric disorders

The importance of metacognition perhaps becomes most apparent when we consider cases where metacognitive monitoring and controlling break down. Metacognitive failures are apparent in commonplace slips such as a student using inefficient study strategies (Bjork et al., 2013) but may become pathological in some psychiatric conditions. Many psychiatric disorders are characterised by denial and unawareness of illness, which can be understood as a failure of metacognition. For instance, patients with Alzheimer's disease are often unaware of their deficits in memory and daily functioning (Green et al., 1993) and patients with schizophrenia may feel like their thoughts do not emanate from themselves but instead that someone else's thoughts are inserted into their minds (Ratcliffe and Wilkinson, 2015).

A range of psychiatric conditions are characterised by distortions in metacognition (Hoven et al., 2019). Abnormalities related to confidence can be identified even in the subclinical or prodromal phases of various psychiatric disorders, and have been shown to be associated with the severity of symptoms. Although the nature of the relationship between psychiatric symptoms and metacognitive dysfunctions remains unclear, deeper insights into the processes underlying metacognition may be valuable in better understanding psychopathology and developing treatment interventions.

Several disorders have been linked with biases in reported confidence. For (sub)clinical OCD, the most consistent finding is an underconfidence bias, especially related to typical OCD contexts, such as checking behaviour (although there are also studies that find no confidence bias in these individuals, e.g. Moritz et al., 2006; Shachar et al., 2013). Two studies exploring confidence in OCD used false biofeedback about

muscle tension as their methodology (Lazarov et al., 2012; Zhang et al., 2017). The studies each conducted two experiments. In the first experiment, participants with low and high obsessive-compulsive tendencies received false biofeedback about muscle tension. In the second experiment, the same procedure was followed comparing a control group and a group that received instructions undermining their confidence. Results demonstrated that participants with high obsessive-compulsive tendencies were more affected by false biofeedback, and this effect was also observed in those who received confidence undermining instructions. This pattern suggests that the reliance of high obsessive-compulsive participants on external cues for assessing internal states could stem from low confidence in their own assessments of internal states. Several studies have also revealed that although OCD patients' memory performance is on par with control participants', the patients tend to exhibit less confidence in their memories (McNally and Kohlbeck, 1993; Moritz and Jaeger, 2018). This underconfidence has been documented in tasks involving both obsessive-compulsive-relevant and neutral content (Karadag et al., 2005). Another study compared episodic memory among three groups: OCD patients with checking symptoms (checkers), OCD patients without checking symptoms (non-checkers), and non-clinical control participants (Macdonald et al., 1997). All three groups performed similarly well on the episodic memory tasks, but the checkers consistently demonstrated less confidence in their memories compared to the other groups.

Another psychiatric disorder that has been associated with an underconfidence bias is depression. In a general knowledge task which typically induces overconfidence, Stone et al. (2001) found that depressed participants exhibited less overconfidence on individual questions compared to control participants and even underconfidence in their overall task performance. Furthermore, Fu et al. (2005) asked depressed patients and controls to provide retrospective self-assessments of performance across four decision tasks: object recognition, general knowledge, social judgement, and

line-length judgement. All participants underestimated their performance, but this underconfidence was more pronounced among the depressed patients. An underconfidence bias in depression was observed in a different domain as well. In an emotion recognition task, Fieker et al. (2016) discovered a negative correlation between self-assessed depression and response confidence in emotion recognition judgements.

Conversely, an overconfidence bias has been associated with clinical and sub-clinical schizophrenia. A study by Warman (2008) indicated a link between delusion proneness and overconfidence. The study employed probabilistic reasoning tasks and discovered that although individuals prone to delusion did not seek less information before making a decision compared to controls, they displayed more confidence in their judgements. Moreover, numerous studies have found that schizophrenia patients exhibit greater confidence, particularly in errors, resulting in a confidence discrimination deficit (e.g., Bhatt et al., 2010; Eifler et al., 2015; Kircher et al., 2007; Malenka et al., 1982; Moritz and Woodward, 2002; Moritz et al., 2005).

Overconfidence bias has been observed to be linked with addiction as well. Two studies focusing on sub-clinical addiction, separated a student population into probable pathological gamblers, problem gamblers, and non-problem gamblers, and utilised a general knowledge task for their experiments (Goodie, 2005; Lakey et al., 2007). In their findings, Goodie (2005) found that pathological gamblers displayed significantly higher confidence, despite having lower task performance compared to the other groups, thereby demonstrating an overconfidence bias. Lakey et al. (2007) similarly found that non-problem gamblers displayed less overconfidence than both the pathological and problem gamblers. Additionally, both studies reported a significant positive correlation between the severity of gambling and overconfidence. Within the realm of clinical addiction, Brewer and Hupfeld (2004) compared the confidence levels of patients with gambling disorder and healthy controls by using

a non-gambling related grammar task. Their findings revealed that despite a lower performance exhibited by the gambling disorder patients, the confidence levels across both groups were similar. Another study conducted by Le Berre et al. (2010) assessed the confidence of patients with alcohol-use disorder using a memory task and a prospective Feeling of Knowing measure. The study showed that these patients had significantly worse memory performance than the healthy controls and were also less accurate in their Feeling of Knowing judgements due to their tendency to overestimate their recognition performance.

The research reviewed above reflects a longstanding interest in whether different psychiatric disorders exhibit overall biases towards high or low confidence. In contrast, there has been less examination of potential alterations in confidence sensitivity, that is, changes in an individual's capacity to discern between correct and incorrect judgements. One key finding in this area is a deficit in metacognitive sensitivity in schizophrenia, a phenomenon primarily driven by overconfidence in errors. Findings on deficits in metacognitive sensitivity in schizophrenia span across several cognitive domains, including memory, visual, and emotional perception (Gawęda et al., 2012; Köther et al., 2012; Moritz et al., 2004; Moritz et al., 2003; Moritz et al., 2012; Peters et al., 2013). Furthering this observation, Davies et al. (2018) discovered that first episode psychosis patients demonstrated significantly lower metacognitive efficiency, measured by meta-d', in a perceptual decision-making task compared to healthy controls. In addition, there is some evidence indicating deficiencies in metacognitive efficiency in other disorders, including addiction (Mintzer and Stitzer, 2002; Moeller et al., 2016; Sadeghi et al., 2017), compulsivity (Hauser et al., 2017), and depression (Hancock, 1996).

The relationship between metacognition and psychopathology has been challenging to formalize due to the significant heterogeneity present within diagnostic categories.

For instance, two patients diagnosed with schizophrenia may exhibit different symptoms (Tsuang et al., 1990). In one patient, positive symptoms such as delusions may dominate, and we may thus expect them to display overconfidence. In contrast, another patient may primarily exhibit negative symptoms like apathy, which should correlate with underconfidence. Moreover, the high comorbidity between disorders, for example between problem gambling and depression (Dowling et al., 2015), further complicates these categorisations and raises questions about the expected cognitive patterns. For instance, should we expect an individual with both addiction and depression to be over- or underconfident?

Such complexities highlight the limitations of employing diagnostic categories, such as in the Diagnostic and Statistical Manual of Mental Disorders (DSM; American Psychiatric Association, 2022) in psychiatric research. There is an increasing agreement within the scientific community that labels like '*schizophrenia*' may not accurately represent singular, biologically grounded, or informative markers of mental health (Huys et al., 2011; Hyman, 2007; Insel et al., 2010). This perspective has catalyzed the development of a field known as computational psychiatry (Friston et al., 2014; Huys et al., 2016; Stephan and Mathys, 2014). This emerging field endeavors to bridge the gap between the fundamental brain processes that govern complex behaviour and the transdiagnostic features that bear significance for mental health. Within the context of this transdiagnostic approach, Rouault, Seow, et al. (2018) found variation in self-reported symptom dimensions to relate to changes in confidence levels in a general population sample. They identified two separate correlations: firstly, a symptom dimension termed Anxiety-Depression (AD) was linked with lowered confidence and enhanced metacognitive efficiency; secondly, a dimension encompassing Compulsive Behaviour and Intrusive Thought (CIT) was associated with elevated confidence and reduced metacognitive efficiency. Yet, these symptom dimensions did not influence the overall accuracy or parameters guiding

decision formation, suggesting a specific link between metacognition and mental health.

Metacognitive abnormalities could be a result or mere byproduct of psychiatric disorders with little or no significance for symptom presentation, but could also be underlying psychiatric symptoms or act in a bidirectional manner, thus constituting a key target for treatment interventions (Wells, 2011). Some causal evidence comes from longitudinal studies looking at symptom severity in relation to the degree to which metacognition is compromised. For instance, Yılmaz et al. (2011) showed that negative metacognitive beliefs prospectively predicted depression and rumination six months later after controlling for initial depression levels. Furthermore, various studies report alterations in metacognition to already appear in sub-clinical at-risk populations suggesting that they may represent early cognitive markers of psychopathology (Eisenacher et al., 2015; Li et al., 2018). Within clinical populations, a correlation between levels of metacognitive abnormalities and symptom severity has been reported. For example, Hancock (1996), found confidence in correct choices to negatively correlate with depression severity and Nicolò et al. (2012) reported metacognitive abilities to correlate with negative symptom severity in schizophrenia. Although experimental research on the relationship between confidence and psychiatric symptoms is scarce, studies on OCD have shown that engaging in compulsive behaviours reduces confidence (Fowle and Boschen, 2011; van den Hout and Kindt, 2003) and, conversely, undermining confidence can foster compulsive tendencies (Cuttler et al., 2013).

The intricate relationship between psychopathology and metacognition warrants further investigation. Given the importance of metacognition for cognitive control, learning, and behavioural adaptation, as reviewed above, the way in which these processes become dysfunctional in psychiatric disorders, requires particular atten-

tion. Understanding these connections can provide a more comprehensive picture of mental health disorders and inform the development of more effective therapeutic strategies.

1.7 Thesis outline

This chapter started out by introducing the concept of metacognition — the monitoring and control of our own cognitive processes, or in simpler terms, thinking about thinking. I have introduced methods used to probe metacognitive processes, for example through confidence judgements, and measures to assess metacognitive performance. Reviewing studies employing diverse neuroimaging techniques, I have discussed the functional localisation as well as our theoretical understanding of metacognition. I have considered whether there are meaningful explicit representations of metacognitive signals, presenting neuroimaging studies that have dissociated metacognitive from object-level processes, as well as studies into the operation of metacognitive signals across tasks encompassing different cognitive domains. Building on these key overall themes in metacognition research, I have explored three particular topics that provide the foundation for the empirical work reported in the following chapters of this thesis. First, I have discussed the functional value of explicit metacognitive signals in guiding adaptive behaviour, such as the ability to effectively seek information, request advice when uncertain, or set reminders based on self-awareness of potential memory lapses. Second, I have introduced the critical role of metacognition in enhancing social interaction and cooperation, for example through relying on others' expression of confidence when seeking their advice or indicating the reliability of one's own opinion within a group setting, as well as effects that the social context can have on self-directed metacognitive appraisals. Third, I have presented findings on metacognitive abnormalities in psychiatric disorders, where both excess overconfidence and underconfidence as well as abnormalities in

metacognitive sensitivity have been shown to be associated with various psychiatric symptoms. The empirical work reported in this thesis builds on the reviewed concepts and methods, and explores the functional role of confidence in flexibly guiding adaptive behaviour.

In the following chapter, Chapter 2, I report two experiments investigating whether confidence can be used to guide behaviour flexibly, depending on the current context. This question is studied by examining how confidence influences information seeking behaviour in relation to the information's value for uncertainty reduction and reward maximisation. In Chapter 3, I describe two experiments that investigate confidence in social settings, dissociating informational and normative social influences on confidence judgements. In Chapter 4, I build on the findings of Chapter 3 by studying neural correlates of confidence under both informational and normative social influences using EEG. In Chapter 5, I examine confidence biases in problem gamblers, and how these biases relate to transdiagnostic symptom dimensions. Finally, in Chapter 6, I summarise the main findings of my empirical work and position these findings within the broader discourse on metacognition.

2

Confidence-Guided Information Seeking

Chapter Abstract

Everyday decision-making is shaped by various uncertainties, leading individuals to experience varying levels of confidence in the accuracy of their choices. Importantly, this sense of confidence can be used for cognitive control, such as seeking information prior to committing to a choice when confidence is low. Whereas most previous studies suggest a fixed relationship between confidence and behaviour, emerging research presents a more flexible, context-dependent perspective on the role of confidence in decision-making. This chapter presents two experimental investigations into the adaptive nature of confidence-guided information seeking behaviour. Experiment 1 assesses the flexibility of the relationship between confidence and information seeking within varied reward contexts. The findings indicate some flexibility whereby, for matched confidence levels, participants seek more information in a task where more reward value is at stake. However, on a trial-by-trial level within a task, participants do not adapt their information seeking behaviour and instead seem to

rely on a heuristic strategy, whereby more information is sought for lower levels of confidence irrespective of reward value. Experiment 2 probes the flexibility of confidence-guided information seeking behaviour by manipulating the likelihood of using the obtained information. Again, findings indicate some flexibility, whereby participants are less inclined to seek information when their likelihood of actually using acquired information would be lower, but on a trial-by-trial level information seeking is heuristically guided by confidence. These findings suggest that although individuals exhibit a degree of flexibility in the way in which confidence guides behaviour, the use of heuristic strategies prevails for more nuanced aspects of the choice environment.

2.1 Introduction

People are faced with numerous decisions on a daily basis. From minor ones like what to eat for breakfast, to more significant ones like where to invest money or who to vote for in an election, decisions can be complicated by various uncertainties. For instance, in selecting a breakfast item, one may face uncertainty about its nutritional content or the quality of its ingredients. Likewise, when casting a vote in an election, uncertainties could emerge from a lack of clarity on a candidate's policy commitments. As a result, people feel varying degrees of confidence in the decisions they make. Generally, individuals are capable of accurately evaluating their own performance and reporting levels of confidence that align well with their objective accuracy (Baranski and Petrusic, 1994; Koriat and Goldsmith, 1996). This ability to assess the quality of one's own decision-making is critical for adaptive behaviour, particularly in situations in which external feedback is unavailable or unreliable. For instance, it allows us to re-evaluate past choices (Folke et al., 2016), make optimal decisions in groups (Bahrami et al., 2010), guide study choices (Hanczakowski et al., 2014) and decisions to set reminders (Risko and Gilbert, 2016), inform trade-offs between exploration and exploitation (Boldt, Blundell, et al., 2019), and direct information seeking choices (Desender et al., 2018; Eliaz and Schotter, 2010).

These examples demonstrate the use of confidence, providing compelling evidence for its functional role in adaptively guiding behaviour. However, much of the potential complexity in this relationship remains to be explored. One such aspect of interest is the flexibility with which confidence is used across varying situations. Previous studies tend to consider a fixed relationship between confidence and subsequent actions, whereby low confidence is associated with one course of action and high confidence with another. For example, several studies have documented a consistent relationship between confidence and information seeking, with lower confidence

driving individuals to seek additional information in order to reduce uncertainty and make more accurate decisions, whereas higher confidence signals that there is no need for further information before committing to a choice (Boldt, Blundell, et al., 2019; Desender, Boldt, et al., 2019; Desender et al., 2018; Desender, Murphy, et al., 2019; Meyniel et al., 2015).

However, this dynamic could be more nuanced in actual practice. Acquiring information can be costly in terms of time and effort. It has been proposed that even the process of accumulating evidence for perceptual decision-making is effortful, such that individuals tend to optimise the time and effort spent on it (Drugowitsch et al., 2012). People are more likely to invest additional effort when they anticipate that such effort will be rewarded and that it will contribute to achieving the expected outcomes (Frömer et al., 2021; Schevernels et al., 2014; Shenhav et al., 2013; Shenhav et al., 2021). The value of additional information depends not only on the marginal increase in the likelihood of making a correct or better decision, but also the marginal increase in value from selecting the correct or better choice. Consequently, low confidence alone is not a sufficient indicator of information need or value. Rather, contextual factors might be expected to influence how confidence guides information seeking choices, which should plausibly reflect calculations of both the likelihood of making a better choice and the value of doing so in the present context, which can then be weighed together against the cost of acquiring further information. The experiments in the present chapter therefore explore the degree to which the relationship between confidence and information seeking is flexibly modulated by the perceived value of additional information.

A recent study by Carlebach and Yeung (2023) on advice-taking discovered that indeed the relationship between confidence and advice seeking behaviour is context-dependent rather than fixed. The study showed that when participants were aware

of an advisor's reliability, they were more inclined to seek advice in situations in which their own confidence was low. This finding suggests that individuals used their level of confidence as a self-monitoring mechanism to recognise when help was needed. However, a reversal of this behaviour was observed when participants did not know the advisor's quality and no objective feedback was provided. In such circumstances, individuals tended to solicit advice when their own confidence was high. Carlebach and Yeung (2023) interpret this seemingly paradoxical behaviour as indicating the strategic use of confidence as a feedback tool for evaluating an advisor's reliability: when individuals are certain about their own decisions, they can also assess an advisor's competence based on whether the advice aligns with their own judgement or not. Assessing an advisor's reliability in this way can be valuable in order to inform choices about whether to follow an advisor's guidance for future decisions. Together, these findings support the idea that confidence may guide behaviour in a flexible manner, enabling individuals to adapt their strategies to the specific context and current goals they are facing in complex and dynamic environments.

Conscious representations like confidence are believed to play a critical role in this flexible and goal-directed guidance of behaviour (Dehaene and Changeux, 2011). Further investigations into the constraints and affordances of the dynamic interaction between metacognitive monitoring and control could yield important insights into the cognitive flexibility that is characteristic of human cognition and enables us to adapt in response to changing demands (Ionescu, 2012). Empirical studies have begun to illuminate the interplay between these two facets of metacognition. For instance, Boldt and Gilbert (2022) implemented a multivariate pattern analysis on fMRI data and found that patterns associated with different levels of metacognitive monitoring, measured as low to high levels of confidence, could differentiate distinct acts of metacognitive control, assessed via the varying desire for setting reminders.

These findings underscore that metacognitive monitoring and control are interlinked and coordinated, with their engagement marked by corresponding patterns in brain activity. Research by Qiu et al. (2018) proposes a functional distinction in the metacognitive network, with different brain regions contributing to metacognitive monitoring and control. Qiu et al. (2018) showed in a decision–re-decision paradigm that dorsal anterior cingulate cortex (dACC) activity during the re-decision phase scaled with decision uncertainty and correlated with individual metacognitive uncertainty monitoring abilities. In contrast, lateral Frontopolar Cortex (lFPC) activity during re-decision processing scaled with decision uncertainty reduction and correlated with individual accuracy changes. The findings suggest that the dACC in particular is pivotal in metacognitive monitoring, whereas the lFPC substantially contributes to metacognitive control. These insights are echoed in a study by Bang et al. (2020) which examined the mapping of private confidence states onto public confidence reports, facilitating the flexible contextual modulation of behaviour in social situations. dACC and pregenual anterior cingulate cortex (pgACC) tracked private confidence — reiterating their role in metacognitive monitoring. In contrast, lFPC activity tracked the need for context-dependent private-public mapping — reiterating its role in metacognitive control. Moreover, lFPC-dACC connectivity varied with task requirements, increasing connectivity when a larger shift in private to public confidence mapping was needed.

Research into metacognitive monitoring and control has primarily been conducted in the domains of perception, memory, and value-based decision-making. Confidence was first studied in the context of psychophysics (Peirce and Jastrow, 1884). The advantage of studying confidence in perceptual tasks lies in the experimenter's ability to have precise control over the evidence accessible to a decision-maker at any given moment (Gold and Shadlen, 2007). Furthermore, a more comprehensive understanding of the neural and computational basis of perceptual judgements, rel-

ative to other domains, facilitates the separation of first-order task judgements from second-order confidence judgements, more easily enabling the examination of their relationship. More recently, the study of confidence has expanded to include value-based choices (De Martino et al., 2013; Lebreton et al., 2015). In general, principles of confidence in decision-making appear to apply similarly in value-based decision-making as in perceptual and memory studies (De Martino et al., 2013), but there are also important differences. In value-based decision-making, it can be challenging to assess first-order accuracy and, consequently, second-order accuracy. For instance, if a person initially prefers an apple over a pear but subsequently changes their mind, it is difficult to determine whether the person made an error or genuinely altered their preference. Furthermore, in the context of value-based decisions, choices often adhere to a gradient of better-to-worse rather than a binary paradigm of right-or-wrong that is typically associated with perceptual decisions. Selecting a slightly less optimal choice in value-based decisions still allows the person to derive some value from the chosen option. For instance, if a person mistakenly chooses an apple, only to later realise that a pear would have been the preferred choice, they still gain the benefit of consuming the apple, which is generally better than no reward at all. In such value-based decisions, seeking additional information when confidence is extremely low may thus not always be the logical course of action. This is because such low confidence could suggest that the options have very similar value, making the slightly superior choice inconsequential and possibly not worth the time and effort spent on discerning which option is marginally better.

This study aims to investigate whether individuals demonstrate flexibility in the relationship between confidence and information seeking, adapting to specific reward contingencies by considering factors like the aforementioned one where the incremental reward from making a better choice should be taken into account when deciding whether low confidence signals the need for additional information. By integrating

a value component into a perceptual decision task, this study effectively sidesteps the difficulty of determining first-order accuracy due to the subjective nature of value, linking value to perceptual stimulus attributes instead. Two experiments were conducted to investigate the interplay between confidence, information seeking behaviour, and task context. The primary goal was to ascertain whether individuals exhibit a fixed relationship between confidence and information seeking, or if they show strategic flexibility in this relationship, adjusting to the particular reward contingencies present.

In Experiment 1, I contrasted the value of information as it pertains to reducing uncertainty with its value in terms of increasing reward, investigating whether individuals adapt their information seeking strategies based on the reward context. The experiment involved participants executing identical tasks under two conditions: one in which rewards were exclusive to correct responses and incorrect responses resulted in losses (as is common in studies examining the link between confidence and information seeking through perceptual decision tasks), and another in which incorrect responses still garnered some, albeit reduced, reward (as is common in real-world value-based decisions). I hypothesised that participants would generally seek less information in the latter scenario: whereas the benefit of information to improve decision quality remains constant between conditions, its value of doing so is reduced in the second scenario. This prediction underscores the idea that the relationship between confidence and information seeking is flexibly modulated by the perceived value of additional information. Additionally, I explored whether participants would also adjust this relationship in more subtle ways, considering the value of the reward at stake not just between conditions, but also on a trial-by-trial basis within a single condition.

Experiment 2 centered on how individuals tailor their information seeking strategies

based on their propensity to utilise the obtained information. Here, I manipulated the cost associated with revising one's decision after acquiring additional information. I hypothesised that as the cost of changes of mind rises, individuals would be less inclined to seek further information, given that the value of the newly acquired information diminishes due to a reduced propensity to act on it. In line with Experiment 1, I thus in Experiment 2 examined whether participants would exhibit strategic flexibility by taking into account the potential value of extra information based on their likelihood of using it.

Through this study, I bridge the domains of value-based decision-making, which assesses how individuals balance costs and benefits in their decisions, and cognitive monitoring and control, which focuses on how individuals monitor and adapt their cognitive processes and behaviours to achieve their objective, in order to investigate the flexibility in the use of metacognitive signals to adaptively guide behaviour.

2.2 Experiment 1

In Experiment 1, I examined the relationship between confidence and information seeking in two distinct contexts, holding constant the inherent value of information for uncertainty reduction while varying the implications of such reduction for potential rewards.

Participants carried out a perceptual decision task in which they judged which of two stimuli appeared more red, with the degree of redness signifying the stimulus value. Two conditions were designed: In one condition, from here on termed the '*Accuracy*' condition, participants were only rewarded for correct choices, receiving points proportional to the value of the accurately selected stimulus (as indicated by its redness). Incorrect choices resulted in a loss of points, again proportional to the chosen stimulus' value. In contrast, in the other condition, from here on

termed the '*Value*' condition, points were awarded for any choice made, but correct decisions yielded more points, reflecting the value of the redder, and hence more valuable, stimulus. The experiment's central prediction was that participants would seek more information in the Accuracy condition compared to the Value condition, reinforcing the notion that the perceived value of additional information can dynamically modulate the relationship between confidence and information seeking behaviours.

Further, I explored whether participants would adjust this relationship in light of the reward value at stake, not only between the two conditions but also within a single condition on a trial-by-trial basis. In the Accuracy condition but not the Value condition, I anticipated an increase in information seeking for particularly red and thus particularly valuable stimuli. In this case, the stakes are high, with significant potential rewards and losses contingent on the correctness of the choice. In contrast, in the Value condition but not the Accuracy condition, I hypothesised that participants would seek reduced information when low in confidence, i.e., when choosing between very similar stimuli, because the difficulty associated with discriminating between the stimuli also means that they offer comparable rewards. Hence, although additional information could decrease uncertainty, the associated increase in reward value would be minimal.

Thus, Experiment 1 investigates whether confidence-guided information seeking is flexibly adjusted based on the reward context, with participants balancing the value of additional information against the costs of obtaining it.

2.2.1 Methods

Participants. Participants were recruited through the online recruitment platform Prolific (www.prolific.co). Experiment 1 was completed by 30 participants. One

participant never opted to seek information and was therefore excluded, leaving a sample of 29 (12 female, mean age ± SD: 34.7 ± 7.4). Participants were compensated with a base amount of £10/hour plus an additional reward depending on performance in the task. The experiment was approved by the University of Oxford's Medical Sciences Interdivisional Research Ethics Committee.

Procedure. Participants performed a series of trials of a perceptual decision task (see Figure 2.1). During the task, they were asked to determine which of the two presented stimuli was more red on average. Each trial began with a fixation cross displayed for 1000 ms, followed by two stimulus displays. Each stimulus consisted of eight coloured shapes arranged in a circle, with a random set of four shapes being masked in the initial presentation. After 500 ms, the stimuli disappeared, and participants were prompted to make a choice and indicate their level of confidence on a visual analogue scale ranging from 50% (guessing) to 100% (completely certain). On 80% of trials, participants had the opportunity to seek additional information before committing to a final choice. If they chose to do so, the stimulus was displayed again for 500 ms, this time revealing all eight coloured shapes that comprised each of the stimuli to be compared. On 15% of trials, participants were not given a choice but had to view the stimulus again, enabling comparison of whether the value of information in terms of increasing accuracy was the same between conditions, and the final 5% of trials were '*catch-trials*' on which participants were immediately moved to the next trial to ensure sensible initial choices.

The mean colour of all eight points, represented as $[C, 0, 1-C]$ in RGB space, could range in 100 steps between pure red ($[1, 0, 0]$) and pure blue ($[0, 0, 1]$). There were three difficulty levels, with the average colour levels of the two stimuli being apart 2 (difficult), 8 (medium), or 14 (easy) steps. The step-size was adjusted during an initial 60 staircasing trials at medium difficulty using a log-adaptive two-down

one-up staircasing procedure (adapted from Rouault, Seow, et al. (2018)), leading to participant accuracy of around 72%. In this way, participants with very precise colour discrimination abilities got to see stimuli in different shades of purple, whereas those with poorer colour discrimination abilities encountered stimuli from a larger range between red and blue.

In contrast to classic perceptual decision tasks, the task used here had a value-based element. Critically, the nature of the value associated with participants' decisions varied across two conditions. The average stimulus colour determined each stimulus' value, ranging from 0 points for the most blue stimulus to 100 points for the most red stimulus, relative to each participant's range of colours, with accumulated points (awarded according to scoring rules described below) determining the participants' monetary bonus at the end of the experiment.

In one condition (Accuracy condition), participants gained the points associated with their stimulus only if they chose the correct (i.e., more red) stimulus, and lost the points associated with their chosen stimulus if they chose the incorrect (i.e., more blue) stimulus. Participants read the following instructions: '*Now you can start earning points in this game. The points will be converted into a monetary reward at the end of the experiment. Points range from 0 for the most blue circles to 100 for the most red circles. If your response is CORRECT (i.e. you choose the more red circle), you will GAIN the points corresponding to the average colour of your chosen circle. If your response is INCORRECT (i.e. you choose the more blue circle), you will LOSE those points instead.*'

In the other condition (Value condition), participants gained the points, i.e. the value, associated with whichever stimulus they chose. Participants read the following instructions: '*Now you can start earning points in this game. The points will be converted into a monetary reward at the end of the experiment. The points you get*

depend on the average colour of the circle you choose. Points range from 0 for the most blue circles to 100 for the most red circles, so you will win more points if you choose the circle with the more red colour patches.'

In both task conditions, on every trial, the points that participants received or lost were displayed in red if they chose the more red option and in blue if they chose the more blue option, providing feedback on their choice. Thus, although in the Value condition, all feedback would signal a win of points, the colour of the feedback still allowed participants to know whether they had made the better or worse choice.

Following the task instructions, participants saw three example trials, detailing how many points they would get for the different choices in trials with different choice options. Participants completed three blocks of 54 trials in one condition, followed by three blocks of 54 trials in the other condition (order counter-balanced across participants).

2.2.2 Results

Prior to running analyses testing my key predictions, I first analysed participants' basic task performance across conditions. Initial choices, made before viewing any additional information, were performed with an average accuracy of 74.4%. An ANOVA was conducted on the choice accuracy, comparing the two conditions (Value versus Accuracy) and three levels of choice option value differences (2 (difficult) versus 8 (medium) versus 14 (easy)). The results indicated a significant effect of choice option value difference on accuracy ($F_{meanDiff}(2, 56) = 52.94, p < .001$). The highest accuracy was observed in the condition with a high choice option value difference, averaging at 90.3% (95% CI = $[82.5, 98.0]$). This was followed by the medium choice option value difference, which showed an average accuracy of 76.1% (95% CI = $[68.2, 83.9]$). The lowest accuracy was found in the trials with a low

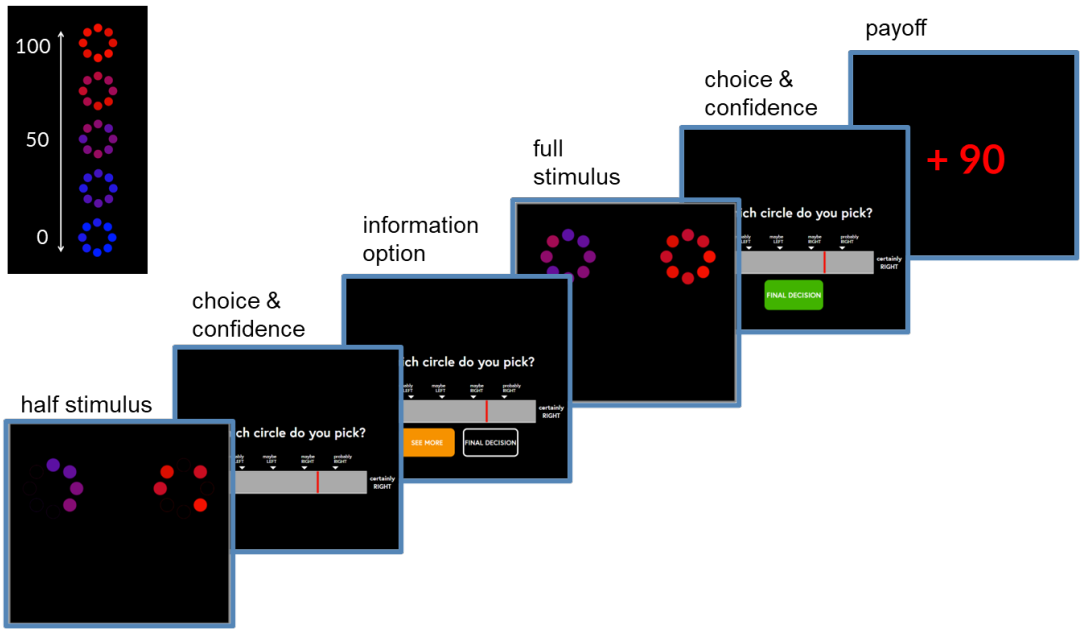


Figure 2.1: Experimental design. Participants had to judge which of two briefly flashed stimuli was more red on average. The value of each stimulus was represented by its colour, with pure blue corresponding to 0 points and pure red equating to 100 points. After viewing the stimuli, participants used a slider to indicate their choice and confidence level. Following this, they were given an option to reveal more of the stimulus and briefly view it again before providing their final choice and confidence rating. The experiment consisted of two conditions: the Value condition and the Accuracy condition. In one condition (Accuracy condition), participants received the point value of their selected stimulus if they correctly identified the more red stimulus, but lost the points if they chose the more blue one. In another condition (Value condition), participants were awarded the point value associated with their chosen stimulus.

choice option value difference, with an average of 57.3% (95% CI = $[50.0, 64.6]$). This is as predicted given that the task is more difficult for smaller differences between choice options. Despite initial expectations that accuracy may vary between conditions because even though the overall perceptual task remains the same, by design there is more at stake for each decision in the Accuracy condition, there was no significant difference in accuracy between the conditions ($F_{condition}(1, 28) = 0.62, p = .43$, $F_{condition \times meanDiff}(2, 56) = 0.07, p = .93$).

Participants reported systematically higher levels of confidence for correct responses compared to incorrect ones as indicated by a 2 (correct versus incorrect) $\times$ 2 (Value versus Accuracy condition) ANOVA ($F_{correct}(1, 28) = 33.49, p < .001$). The lack of an interaction effect between choice-accuracy and condition indicates that there was no consistent difference in this effect across conditions ($F_{correct \times condition}(1, 28) = 2.13, p = .16$). Despite the matched accuracy between conditions, and against expectations, participants reported significantly higher confidence in the Value compared to the Accuracy condition ($F_{condition}(1, 28) = 27.24, p < .001$). Participants' metacognitive sensitivity was found to be greater than chance, as evidenced by area under the ROC curve (AUROC) values significantly greater than 0.5 for both the Value condition ($M = 0.70$, $SE = 0.09$, $t(28) = 11.38, p < .001$) and the Accuracy Condition ($M = 0.69$, $SE = 0.10$, $t(28) = 10.08, p < .001$), with no significant difference between conditions ($t(28) = 0.29, p = .774$)

Across conditions, when given the option to do so, participants on average opted to see more information on 40.4% of the trials. As in previous experiments, the frequency of seeking information varied widely among participants, here ranging from 0.4% to 81.7%. On trials in which participants actively opted to see more information, their first choice accuracy was significantly lower at 64.0% compared to their first choice accuracy on trials on which participants were forced to do so

at 74.6% ($t(28) = 3.43, p < .01$), as was their confidence with $M = 66.5$ compared to $M = 71.6$ ($t(28) = 4.96, p < .001$). When participants actively chose to gather additional information, their accuracy in the task increased significantly from 64.0% accuracy on first choices to 81.3% accuracy on second choices ($t(28) = 5.0, p < .001$), as did their reported confidence from $M = 66.5$ on first choices to $M = 74.1$ on second choices ($t(28) = 9.20, p < .001$).

As a first critical test of the sensitivity of participants' information seeking behaviour to task context, I compared information seeking between the Accuracy and Value conditions. This analysis revealed, as predicted, that participants sought more information in the Accuracy condition than in the Value condition ($t(28) = 3.08, p < .01$). Because participants reported higher confidence for initial choices in the Value condition at $M = 73.4$ compared to the Accuracy condition at $M = 70.0$, and higher confidence is typically associated with reduced information seeking, I conducted a linear mixed-effects model to test if the difference in information seeking between conditions remained after adjusting for reported confidence. Trial-wise information seeking choices were predicted in a logistic regression by reported confidence and task condition, allowing for random intercepts and slopes by participant. The regression model showed that task condition influenced information seeking ($\chi^2 = 6.08, p < .05$) beyond confidence ($\chi^2 = 38.32, p < .001$), with participants seeking more information in the Accuracy condition compared to the Value condition as can be seen in Figure 2.2.

Next, I examined how choice option value difference and average absolute value of choice options impacted both confidence levels and information seeking behaviour. The aim was to determine if people displayed flexibility in their information seeking in more nuanced ways on a trial-by-trial level. I expected that both the choice option value difference and the overall value of choice options would have similar

effects on confidence under both conditions. This is because the overall task is the same between conditions — confidence should be lower for lower choice option value differences because the task becomes more difficult, and confidence should not be affected by overall value of choice options because this does not affect task difficulty. However, any flexibility in how confidence then guides information seeking behaviour by taking into account reward values at stake, would be revealed through an interaction effect between these factors and the condition on information seeking choices.

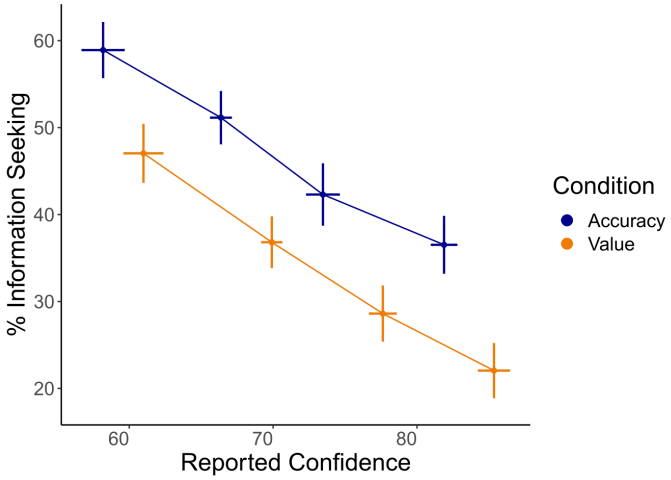


Figure 2.2: Information seeking as a function of reported confidence in the Accuracy and the Value condition on trials in which participants could choose to seek more information. The analysis reported in the main text was performed on single-trial data, but for the purposes of visualisation here, trials were sorted into bins according to quartiles of the distribution of confidence for each participant separately, and the proportion of information seeking choices within each confidence quartile was computed. The graph displays the grand mean across participants and standard error of the mean for both reported confidence and information seeking.

First, I assessed the choice option value difference. I hypothesised that although a lower choice option value difference would lead to lower confidence in both conditions as this makes the task more difficult, only in the Accuracy condition, but not the Value condition, a low choice option value difference would also lead to increased

information seeking. This is because in the Value condition, a low choice option value difference implies that both stimuli offer comparable rewards. Although additional information could decrease uncertainty arising from the low choice option value difference, the increase in reward value from improving accuracy would be minimal. Due to non-convergence of the regression models when including interaction terms in the random effects structure, I specified the models as:

$$\text{confidence} \sim \text{choice option value difference} : \text{condition} + (1 + \text{choice option value difference} + \text{condition} \mid \text{participant})$$

and

$$\text{information sought} \sim \text{choice option value difference} : \text{condition} + (1 + \text{choice option value difference} + \text{condition} \mid \text{participant})$$

The regression on confidence for initial choices revealed significant effects of condition ($\chi^2 = 24.59, p < .001$), with higher confidence in the Value condition compared to the Accuracy condition, reiterating the surprising results reported above. Moreover, as expected, there was a significant effect of choice option value difference on confidence ($\chi^2 = 39.63, p < .001$), with higher confidence for larger choice option value differences, and no interaction effect of condition and choice option value difference ($\chi^2 = 0.24, p = .62$) on confidence. The same pattern emerged in the subsequent logistic regression on trial-wise information seeking choices. There was a significant effect of condition ($\chi^2 = 9.4, p < .01$) on information seeking, reiterating evidence for the primary hypothesis, and a main effect of choice option value difference ($\chi^2 = 26.44, p < .001$), reiterating that in general, more information is sought for more difficult choices. However, against expectations, I did not find a significant

interaction effect between these factors ($\chi^2 = 3.38, p = .07$) on information seeking (see Figure 2.3). Not only was the interaction not significant, but the trend went even in the opposite direction of my predictions, with the slope for information seeking being numerically steeper in the Value as opposed to the Accuracy condition. This suggests that participants did not take into account the fact that in the Value condition, in trials with small choice option value difference, despite low confidence, additional information has low value given that the increase in reward from making the correct/better choice is low.

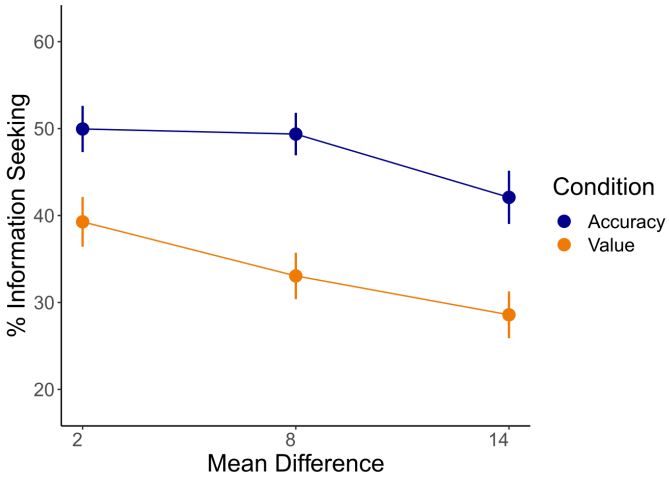


Figure 2.3: Information seeking as a function of choice option value difference in the Accuracy and the Value condition on free choice trials. The graph displays the grand mean across participants and standard error of the mean.

Next, I explored the influence of average absolute stimulus value of choice options. I expected that although a higher average absolute stimulus value of choice options would not lead to increased or decreased confidence in either condition, as this does not affect task difficulty, only in the Accuracy condition, but not the Value condition, a higher average absolute stimulus value of choice options would lead to increased information seeking. This is because in the Accuracy condition a high average absolute stimulus value of choice options implies that the stakes are high, with

significant potential rewards and losses contingent on the correctness of the choice. In contrast, in the Value condition, the stakes are only dependent on the relative value difference between choice options. To ensure model convergence, absolute value was divided into octiles. Paralleling the previous analyses, regression models were specified as:

$$\text{confidence} \sim \text{average absolute stimulus value} : \text{condition} + (1 + \text{average absolute stimulus value} + \text{condition} \mid \text{participant})$$

and

$$\text{information sought} \sim \text{average absolute stimulus value} : \text{condition} + (1 + \text{average absolute stimulus value} + \text{condition} \mid \text{participant})$$

The confidence regression reiterated a significant effect of condition ($\chi^2 = 25.18, p < .001$). Additionally, a significant effect of absolute stimulus value ($\chi^2 = 55.24, p < .001$) on confidence was observed, with participants reporting higher confidence for initial choices on trials with higher average absolute value choice options (see Figure 2.4). Moreover, an interaction effect between condition and stimulus value ($\chi^2 = 9.44, p < .01$) was found, indicating a more pronounced increase in confidence with increasing stimulus value in the Accuracy condition. In contrast, the subsequent logistic regression on information seeking choices demonstrated a significant effect of condition ($\chi^2 = 9.97, p < .01$) on information seeking and no effect of average absolute stimulus value on information seeking ($\chi^2 = 1.40, p = .24$). Against expectations, there was no interaction effect between the condition and average absolute stimulus value on information seeking ($\chi^2 = 0.14, p = .71$; see Figure 2.5). The lack of an interaction effect, underlying which is a lack of the expected

effect of average absolute value on information seeking in the Accuracy condition, is surprising given that there is more at stake for higher value stimuli in this condition. Not only was the interaction not significant, but the trend went even in the opposite direction of my predictions, with the correlation between average absolute stimulus value and information seeking averaged across participants being negative, albeit non-significant, within the Accuracy condition ($r = -.447, p = .27$).

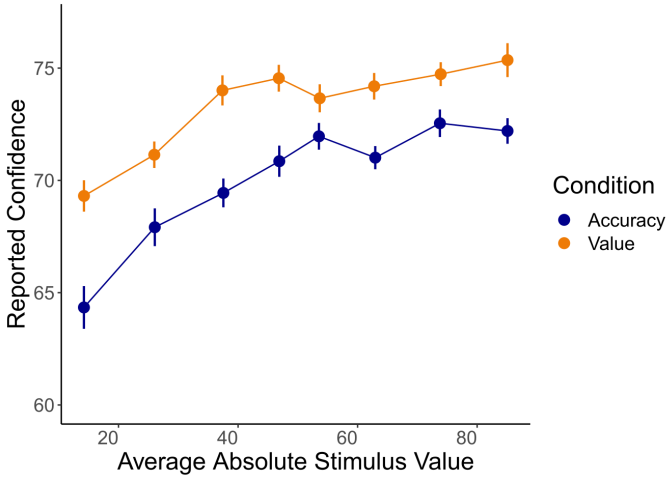


Figure 2.4: Confidence as a function of average absolute stimulus value in the Accuracy and the Value condition on free choice trials. Average absolute stimulus value was divided into octiles. The graph displays the grand mean across participants and standard error of the mean.

To check whether any differences in confidence and information seeking might be influenced by the care taken in initial choices, I next analysed reaction times as an index of this effort. A key question here was whether participants deliberated longer, particularly in the Accuracy condition, when more reward was at stake. However, against this intuition, reaction times for first-order decisions (after eliminating outlier trials for which reaction time had a by-participant absolute z-score greater than 2) did not differ significantly between conditions, with $M = 1009$ ms for the Accuracy condition and $M = 1005$ ms for the Value condition ($t(28) = 0.16, p = .87$). A linear mixed effects model predicting reaction times by choice option value dif-

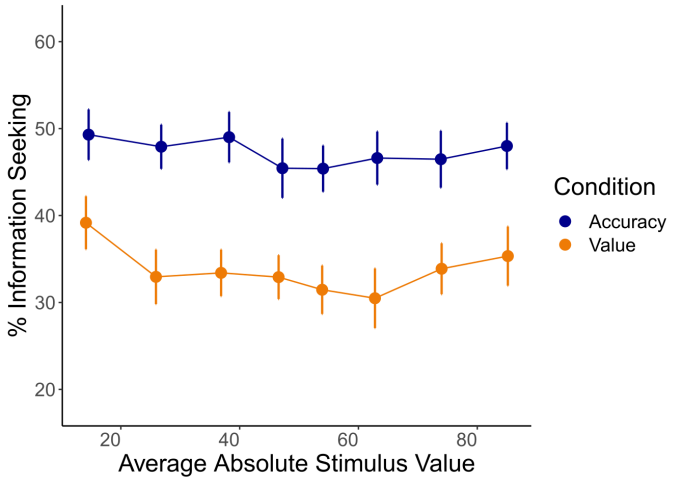


Figure 2.5: Information seeking as a function of average absolute stimulus value in the Accuracy and the Value condition on free choice trials. Average absolute stimulus value was divided into octiles. The graph displays the grand mean across participants and standard error of the mean.

ference and condition and their interaction only showed a significant effect of choice option value difference, whereby participants responded faster on trials with a larger difference between stimuli (i.e., easier trials; $\chi^2_{meanDiff} = 5.11, p < .05$; $\chi^2_{condition} = 0.05, p = .81$, $\chi^2_{condition \times meanDiff} = 0.05, p = .82$). Absolute stimulus value did not significantly affect reaction times in either condition ($\chi^2_{value} = 1.74, p = .19$; $\chi^2_{condition} = 0.01, p = .94$, $\chi^2_{condition \times value} = 0.09, p = .76$).

2.2.3 Discussion

Experiment 1 investigated the flexibility of the relationship between confidence and information seeking behaviour in tasks with differing contingencies between the accuracy of decisions and resulting outcome values. The results provided nuanced support for the hypothesis that the interplay between confidence and information seeking behaviour is flexible and context-dependent. Participants in the experiment exhibited an amplified tendency towards information seeking when their incorrect decisions led to an active loss of rewards, in comparison to a scenario in which incorrect answers simply resulted in a lesser reward. This behaviour was observed

despite the fact that the informative value of the information in terms of uncertainty reduction was equal in both scenarios. Moreover the divergence in behaviour between conditions was evident despite equal levels of accuracy and controlling for participants' self-reported confidence. This finding indicates that the relevance of information was perceived to be heightened when incorrect decisions led to a loss given the more substantial discrepancy in reward between choice options in this case (for instance, a loss of 60 compared to a win of 74, as opposed to a win of 60 versus a win of 74). Although the level of confidence plays a role in one's propensity to seek out information (Boldt, Blundell, et al., 2019; Desender, Boldt, et al., 2019; Desender et al., 2018; Desender, Murphy, et al., 2019; Meyniel et al., 2015), evidently, the context in which the decision-making occurs can alter this relationship (Carlebach and Yeung, 2023). Thus, my findings underscore that the association between confidence and information seeking is not a static one; instead, it is adaptive and sensitive to the particular context. This observation is in line with theories proposing that conscious representations like confidence play a key role in guiding behaviour in a flexible and goal-directed manner (Dehaene and Changeux, 2011).

Although overall there is evidence of flexibility in the relationship between confidence and information seeking, as seen by the difference in information seeking between conditions, this flexibility appears to have its limitations. In particular, I did not find evidence to support a more nuanced prediction about how information seeking would vary as a function of choice option values on a trial-by-trial level. I predicted that a decline in confidence due to smaller choice option value differences should not result in increased information seeking in a scenario in which both options yield a reward (Value condition) as opposed to a scenario in which incorrect choices lead to losses (Accuracy condition). This is because in the former situations, smaller differences suggest that the choice options hold more similar worth, rendering the marginal gain from selecting the correct/better option minimal. However, counter

to these expectations, participants sought more information for smaller choice option value differences in both conditions, instead of altering their strategy between the conditions, whereby in one case, low confidence due to a small choice option value difference indicates that more information is rather unnecessary for reward maximisation, whereas in the other case, low confidence caused by a small choice option value difference does signal the need for more information. The finding that participants did not adjust their trial-by-trial strategy between conditions suggests that participants may have adhered to a heuristic strategy whereby low confidence invariably prompts the need for more information, rather than being flexible in the relationship between confidence and information seeking depending on the context.

I further predicted that a higher average absolute value of choice options would, despite not affecting confidence, lead to more information seeking in the scenario in which incorrect choices lead to losses. This prediction is based on the notion that the potential gain or loss from selecting the correct choice is higher for decisions involving stimuli of higher average absolute value since getting the choice wrong results in losing a significant value, whereas getting it right yields a substantial gain. In contrast, when both choices simply lead to winning the value associated with the chosen option, then the marginal gain from selecting the correct/better option only depends on the relative difference between choice options but is independent of the absolute value of the options. Yet, against these expectations, participants did not modify their information seeking behaviour based on the average absolute value of choice options. This absence of an effect, particularly the lack of distinction between conditions, provides additional support for potential reliance on heuristic strategies on a trial-by-trial basis, hinting at a possible limitation in the context-dependent adaptability of the relationship between confidence and information seeking.

Though one might contend that the absence of condition differences in how choice

option value difference and the average absolute value of choice options impacted information seeking on a trial-by-trial basis could stem from insufficient power in the study, the evidence suggests otherwise. Qualitatively, the anticipated effects should have been prominent, yet were notably absent. Additionally, not only did the expected effects fail to manifest in a significant manner, but the numerical trends also leaned slightly contrary to my theoretical predictions. This clear deviation from the theoretical predictions renders a mere lack of statistical power an unlikely explanation.

Although the primary focus of this study was how task context influenced the relationship between confidence and information seeking, the results revealed unexpected influences of task context on confidence itself. The results show that despite matched accuracy and despite objective feedback provided on every trial, participants were more confident for higher value stimuli. This phenomenon suggests that their confidence reports may at least partially be based on the expected quality of the outcome, aligning with the idea that the calculations of confidence and value are intrinsically linked (Barron et al., 2015; De Martino et al., 2013; Lebreton et al., 2015). This is further reflected in participants' generally higher confidence in the scenario where all choices lead to some win, compared to the scenario with potential losses, as they gain more reward overall in the former.

Taken together, the findings from Experiment 1 suggest that although individuals exhibit a degree of flexibility in their information seeking behaviour in relation to the context, as highlighted by the variation between conditions, they seem to lean on heuristic strategies for nuanced decision-making processes. This is underscored by the lack of condition differences in how choice option value differences and average absolute stimulus values influenced information seeking behaviour on a trial-by-trial basis. These findings propose some flexibility but also hint at the potential

boundaries of this adaptability, suggesting a possible inclination of individuals to adopt broad-brush strategies that might not be sensitive to nuanced contextual factors.

In the following experiment, I aimed to further explore the flexibility of confidence-guided information seeking behaviour, and the potential limitations thereof, particularly by examining whether individuals modify their information seeking strategies based on the likelihood of actually using the acquired information.

2.3 Experiment 2

Experiment 2 was devised to investigate whether individuals modify their information seeking strategies contingent on their propensity to utilise the gathered information. I postulated that if the likelihood of revising an initial decision decreases, participants' likelihood to pursue further information would proportionally diminish. The underpinning premise is that the utility and thereby the value of new information declines when there is a decreased inclination to act upon it.

As in Experiment 1, the task required participants to distinguish the redder of two stimuli, with the colour of the stimulus translating into its value. Participants won the value tied to their selected stimulus for correct choices, but lost this value for incorrect choices, mirroring the '*Accuracy*' condition from Experiment 1. A significant shift in this experiment was the incorporation of a cost for revising one's decision after viewing additional information. This cost varied across different blocks and was deducted from the participant's total points that contributed to their monetary reward. Participants were briefed on this cost at the onset of each block, and upon any attempted revision to their initial decision.

My hypothesis suggested that an increase in the cost of changing one's decision

would make participants less likely to seek additional information, due to the reduced value of new information based on the decreased propensity to act on it. Notably, the task itself was consistent across the blocks that varied the cost of revising an initial choice, such that there should not be any variance in confidence between the different cost conditions. Therefore, any shift in information seeking behaviour between cost conditions would be independent of confidence and thus imply a degree of flexibility in the relationship between confidence and information seeking.

The principles guiding Experiment 2 resonate with the economic concept of rational inattention (Sims, 2003). According to rational inattention, individuals optimise the allocation of their finite cognitive resources, selectively attending to the most pertinent information and disregarding the rest. Similarly, in this experiment, participants were hypothesised to adjust their information seeking based on the perceived value of additional information and the cost of modifying their decision. The potential cognitive and monetary cost of altering their decision after receiving new information might lead participants to be '*rationally inattentive*', choosing not to seek additional information that they are less likely to act upon. For example, a real-world analogy that encapsulates this idea could be the act of purchasing a non-refundable airline ticket. In this case, the consumer makes an initial decision based on the available information about prices, schedules, and other travel needs. Once the non-refundable ticket is purchased, the cost of changing the decision (i.e., altering or cancelling the flight) is quite high. So, the consumer is less likely to seek more information about other flight options, as they are less likely to act on it given the high cost of changing their decision. Just as in the experiment, individuals in real-world scenarios must strategically allocate their resources, such as time, attention, or money, to maximise their overall utility.

2.3.1 Methods

Participants. I recruited participants for Experiment 2 via the online platform Prolific. A total of 30 participants completed the experiment. Two participants never opted to seek information and were therefore excluded, leaving a sample of 28 (10 females, average age ± SD: 34.5 ± 7.6). Participants received a base compensation of £10/hour, along with an additional reward based on their task performance. The University of Oxford's Medical Sciences Interdivisional Research Ethics Committee approved the experiment.

Procedure. Participants performed a series of trials of the same perceptual decision task as in Experiment 1 (see Figure 2.1). In the task, participants had to determine which of the two presented stimuli was more red on average, with the stimulus colour defining its worth. As in Experiment 1, there were three difficulty levels, with the average colour levels of the two stimuli being apart 2 (difficult), 8 (medium), or 14 (easy) steps. On all trials, participants only gained the points associated with their chosen stimulus if they correctly identified the more red stimulus, and lost the points if they chose the incorrect (more blue) stimulus (i.e., equivalent to Experiment 1's Accuracy condition). As in Experiment 1, on 80% of trials, participants had the option to seek additional information before committing to a final choice. If they chose to do so, the stimulus was displayed again for 500 ms, revealing all eight colour points. On 15% of trials, participants were required to view the stimulus again without a choice, and the remaining 5% of trials were '*catch-trials*' on which participants were immediately moved to the next trial to ensure sensible initial choices. After each choice, feedback was provided on the value of the chosen outcome.

The key manipulation in this experiment was the introduction of a cost for changing one's mind after seeing additional information. Participants were informed of this

cost at the start of every block. Additionally, when participants moved their cursor on the choice and confidence slider to the opposite (compared to their initial choice) side during their second choice, a message appeared, reiterating the cost of changing their mind. There were three cost conditions: 5 points, 15 points, or 30 points deducted from the points contributing to the participant's monetary performance reward. Participants read the following instructions: '*When you click the "see more" button, and after seeing all 8 colour patches you decide that you want change your mind (e.g. on your first choice you thought that the left circle looks more red on average but after seeing more you now think that actually the right circle is more red), you will have to pay in points which make up your monetary reward.*'

After 60 staircasing trials, participants went through 30 practice trials to familiarise themselves with indicating their confidence in the initial choice, seeking additional information, and understanding the cost implications of altering their initial choice. During these practice trials, no actual points were lost or won. Subsequently, participants completed six blocks of 54 trials each, with two blocks per cost condition in random order.

2.3.2 Results

Participants achieved a mean accuracy of 75.8%. An ANOVA with a 3×3 design was conducted to examine the effects of the change-of-mind cost (5 points versus 15 points versus 30 points) and the choice option value difference (2 (difficult) versus 8 (medium) versus 14 (easy)). This analysis revealed a significant effect of the choice option value difference on accuracy ($F(2, 54) = 82.64, p < .001$), but no significant differences in accuracy across the various change-of-mind cost conditions nor a significant interaction between the cost condition and the choice option value difference ($F_{cost}(2, 56) = 0.05, p = .96$, $F_{cost \times meanDiff}(4, 108) = 0.38, p = .82$). These findings suggest that increased costs associated with changing decisions did not

compel participants to exercise more caution in making their initial choices. The accuracy rates were approximately equivalent across the 5-point cost (76.0%), 15-point cost (76.1%), and 30-point cost (75.2%) blocks.

Participants consistently reported higher confidence for correct responses compared to incorrect ones ($t(27) = 8.27, p < .001$). Moreover, their metacognitive sensitivity was significantly above chance, as demonstrated by AUROC values significantly greater than 0.5 ($M = 0.69$, $SE = 0.07$, $t(27) = 14.20, p < .001$). An ANOVA indicated no significant difference in reported confidence between change-of-mind cost conditions ($M = 69.6$, $M = 69.9$ and $M = 68.6$, in the 5-, 15- and 30-point change-of-mind cost blocks, respectively, $F(2, 54) = 0.38, p = .68$). Thus, unlike in Experiment 1, the primary manipulation in this experiment did not influence confidence judgements.

When given the option, participants opted to seek more information on an average of 30.4% of trials. However, there was considerable variation among participants, with information seeking frequency ranging from as low as 1.3% to as high as 72.6%. On trials in which participants decided to gather additional information, task accuracy improved significantly from an initial 62.0% to a subsequent 79.0% ($t(27) = 5.88, p < .001$). Alongside this, a significant increase in reported confidence was observed, from an initial $M = 63.46$ to a subsequent $M = 68.99$ ($t(27) = 7.04, p < .001$).

As a key check of the effectiveness of the cost manipulation, I compared participants' likelihood of changing their mind from their initial to their final decisions across blocks, focusing on the trials on which participants were forced to view the stimulus again. Looking at this subset of trials allows an analysis that is not confounded with the likelihood of requesting further information in the first place. An ANOVA revealed a significant effect of the cost condition on participants' likelihood of changing their minds ($F(2, 54) = 4.63, p < .05$). In these forced information trials,

participants altered their initial decisions 22.9% of the time in the 5-point condition, 17.5% in the 15-point condition, and 17.4% in the 30-point condition. These trends generally align with the intended effect of the manipulation, although Figure 2.6 reveals substantial inter-individual variation, with some participants paradoxically more inclined to change their minds in higher cost conditions, contrary to expectations. Notably, the divergence here is a higher change-of-mind likelihood for the 5-point cost condition in comparison to the 15- and 30-point conditions, which in turn are similar. Follow-up pairwise comparisons showed that only the 5-point and 15-point conditions differed significantly ($p < .01$ for 5 versus 15, $p = .166$ for 5 versus 30, $p = .796$ for 15 versus 30).

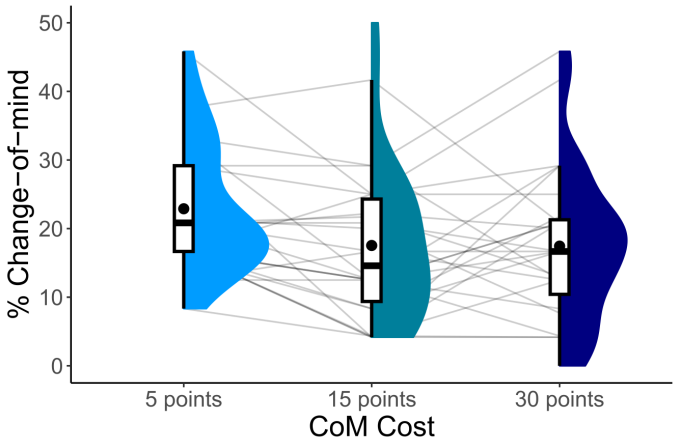


Figure 2.6: Change-of-mind frequency in the three change-of-mind (CoM) cost conditions on forced view again trials. Faint lines show individual participant means. Violin and box plots show the distributions of participant means.

I then investigated if participants took this difference in information-use likelihood into account when seeking information during trials on which they could freely choose whether or not to seek information. As expected, participants sought the most information in the 5-point condition (32.3%), followed by the 15-point condition (31.7%), and the least in the 30-point condition (27.1%; see Figure 2.7). A logistic linear mixed-effects model was used to predict information seeking choices

based on the change-of-mind cost condition and adjusted for reported confidence, allowing for random intercepts and slopes per participant. As expected, the model revealed significant effects of both confidence ($\chi^2 = 79.11, p < .01$) and the change-of-mind cost condition ($\chi^2 = 6.46, p < .05$) on information seeking behaviour (see Figure 2.8). Notably, contrasting the change-of-mind propensities reported earlier, the difference here lies in a lower likelihood of seeking information in the 30-point cost condition compared to the now similar 5- and 15-point cost conditions. As shown in Figure 2.8, this discrepancy in information seeking likelihood exists across all confidence levels, excluding the highest, which likely mirrors a floor effect in information seeking choices. Thus, overall, the results provide further evidence for the flexibility in confidence-guided information seeking behaviour.

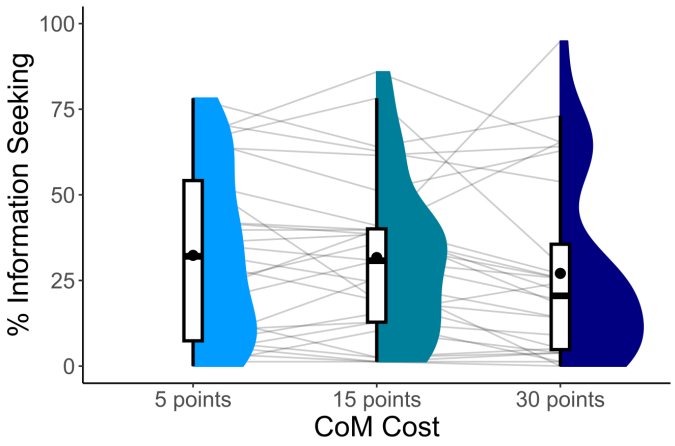


Figure 2.7: information seeking in the three change-of-mind (CoM) cost conditions on free choice trials. Faint lines show individual participant means. Violin and box plots show the distributions of participant means.

As in Experiment 1, I conducted further analyses to examine the flexibility of participants' information seeking choices on a trial-by-trial basis. Given that a higher average absolute stimulus value denoted higher stakes — with potentially significant rewards or losses hinging on the accuracy of participants' decisions — I anticipated a consequent increase in information seeking behaviour. However, this

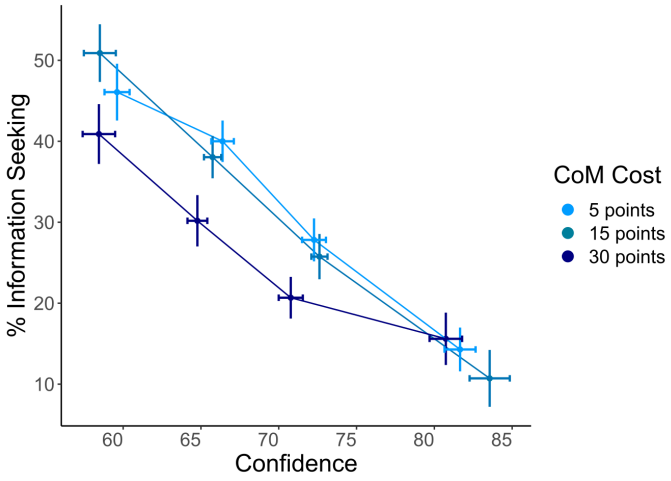


Figure 2.8: Information seeking as a function of reported confidence in the three change-of-mind (CoM) cost conditions on free choice trials. For each participant, confidence was divided into quartiles and the proportion of information seeking choices within each confidence quartile was computed. The graph displays the grand mean across participants and standard error of the mean for both reported confidence and information seeking.

hypothesis was not substantiated by a logistic regression analysis, which showed no significant increase in information seeking for trials with a higher average absolute stimulus value of choice options ($\chi^2_{value} = 2.39, p = .12$, $\chi^2_{cost} = 4.48, p < .05$, $\chi^2_{cost \times value} = 0.98, p = .32$; see Figure 2.9). Instead, the data hinted at a U-shaped pattern, showing a propensity for increased information seeking when average absolute stimulus value of choice options was at either the low or high extreme. This trend can be explained by the fact that extreme average absolute stimulus values are only possible when both options carry very low or very high values, thus resulting in choices that are inherently difficult due to the close proximity of these values. When the average absolute stimulus value was examined in relation to confidence, the findings echoed those observed in Experiment 1. A linear mixed effects model predicting confidence reports by average absolute stimulus value and condition found a significant impact of stimulus value on confidence, with higher confidence for choices between higher value options ($\chi^2_{value} = 23.45, p < .001$, $\chi^2_{cost} = 3.42, p = .06$, $\chi^2_{cost \times value} = 1.52, p = .22$; see Figure 2.10).

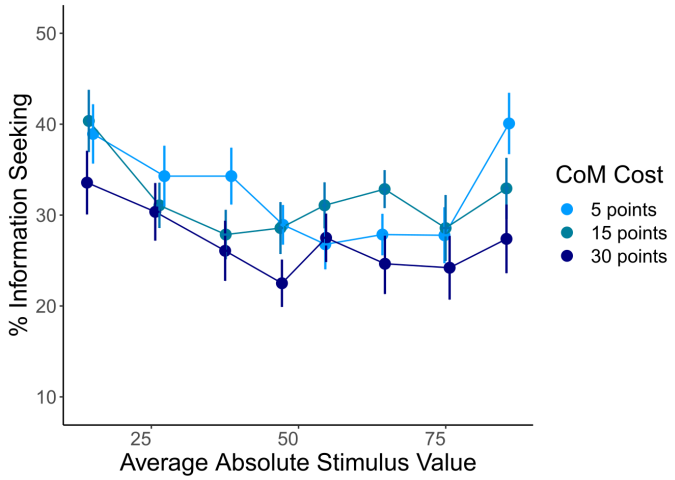


Figure 2.9: Information seeking as a function of average absolute stimulus value in the three change-of-mind (CoM) cost conditions on free choice trials. Average absolute stimulus value was divided into octiles. The graph displays the grand mean across participants and standard error of the mean.

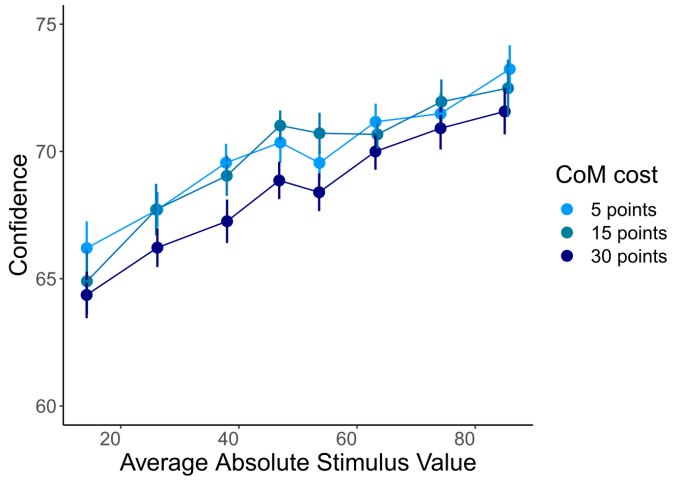


Figure 2.10: Confidence as a function of average absolute stimulus value in the three change-of-mind (CoM) cost conditions on free choice trials. Average absolute stimulus value was divided into octiles. The graph displays the grand mean across participants and standard error of the mean.

2.3.3 Discussion

Experiment 2 examined the flexibility of confidence-guided information seeking behaviour, investigating whether individuals adjust their information seeking strategies according to the probability of utilising the additional information. By manipulating the cost of changes of mind, the experiment altered the likelihood of participants using the newly-acquired information. On trials in which additional information was compulsory, participants demonstrated a higher propensity to change their initial decisions in the scenario with the lowest change-of-mind cost (5 points) compared to those with medium (15 points) and high (30 points) costs. A critical question stemming from these findings was whether this cost-dependent likelihood of changing an initial decision would in turn impact the inclination to seek information in the first place. Indeed, participants were least likely to seek additional information in the high cost condition (30 points), in contrast to the medium (15 points) and low cost (5 points) conditions. This pattern was evident despite the potential informative value of the information in terms of reducing uncertainty remaining constant across conditions. The observed divergence of participants' behaviour between cost conditions was evident despite matched accuracy levels for initial choices and controlling for self-reported confidence levels. This finding suggests that when the likelihood of acting on additional information decreased due to a high change-of-mind cost, the perceived value of said information correspondingly diminished. These results echo findings from Experiment 1, reiterating that confidence is a significant factor in one's propensity to seek information (Boldt, Blundell, et al., 2019; Desender, Boldt, et al., 2019; Desender et al., 2018; Desender, Murphy, et al., 2019; Meyniel et al., 2015), but at the same time underscoring that the context of decision-making can affect this relationship (Carlebach and Yeung, 2023), further highlighting that the relationship between confidence and information seeking is flexible and context-dependent, rather than static.

Although the findings suggest a level of flexibility in the relationship between confidence and information seeking, this adaptability appears to have its limits. I found no support for a more nuanced prediction on a trial-by-trial basis about how information seeking would change based on the average absolute value of choice options. I hypothesised that a higher average absolute value of choice options would lead to increased information seeking, despite having no impact on confidence. This premise stemmed from the idea that the potential gains or losses associated with making the correct decision become more significant when decisions involve higher valued stimuli. Against this prediction, participants did not adjust their information seeking behaviour based on the average absolute value of the choice options, echoing the findings of Experiment 1 and suggesting a potential reliance on heuristic strategies. Furthermore, also replicating findings from Experiment 1, the study uncovered unexpected influences of task context on confidence itself. The data showed that despite consistent accuracy levels and objective feedback provided on every trial, participants displayed higher confidence for choices involving higher value options.

2.4 General Discussion

The findings of the two experiments reported in this chapter provide converging support for the idea that information seeking behaviour is context-dependent and strategically flexible. In Experiment 1, participants adjusted their information seeking behaviour based on the potential implications of their decisions, seeking more information when an incorrect choice led to a loss of rewards, rather than merely receiving lesser rewards. However, the data also suggests that despite the evident context-dependence, this flexibility in information seeking was limited. Specifically, participants did not adjust their strategies based on subtler details of the decision context, such as the average absolute value of choice options and differences in choice option values on a trial-by-trial basis, suggesting a heuristic-driven decision-making

process. Experiment 2 built upon these findings from Experiment 1 by examining the effect of decision revision cost on information seeking. The results reinforced the context-dependent flexibility observed in Experiment 1 but also mirrored similar limitations. Participants adjusted their information seeking based on the cost of changing their minds, showing sensitivity to the likelihood of utilising potential new information. Yet, this behavioural flexibility did not extend to more nuanced aspects of the decision-making context on a trial-by-trial basis, as evidenced by a lack of consideration for the average absolute value of choice options in information seeking choices. Taken together, it appears that although individuals display a degree of flexibility in their confidence-guided information seeking behaviours depending on the context, they may default to heuristic strategies for nuanced decision-making aspects.

The observations from these two studies bolster the proposition that confidence not only guides information seeking in a fixed manner, whereby lower confidence signals a need for more information and higher confidence signals no need for further information before committing to a choice, as established by past research (Boldt, Blundell, et al., 2019; Desender, Boldt, et al., 2019; Desender et al., 2018; Desender, Murphy, et al., 2019; Meyniel et al., 2015), but that this relationship between confidence and information seeking adapts flexibly to the context (Carlebach and Yeung, 2023), such as potential consequences of decisions and the cost of decision revision.

The observed limitations in considering subtle factors beyond confidence in determining information seeking choices underscore the close link between metacognitive monitoring and control (Nelson and Narens, 1990) and the causal influence of confidence on information seeking. For example, Desender et al. (2018) utilised a perceptual manipulation to induce different levels of confidence across conditions while maintaining participants' objective performance constant. In the lower confidence

condition, participants were more likely to seek additional information, providing critical causal evidence for the role of confidence in information seeking. In memory research, artificially boosting individuals' confidence during word pair learning made them less likely to choose to study those pairs again, even though their performance remained unchanged (Metcalfe and Finn, 2008).

This pronounced influence of confidence on information seeking aligns with the idea that confidence signals are automatically encoded during decision-making processes (Gherman and Philiastides, 2015; Kepecs et al., 2008; Kiani and Shadlen, 2009; Lebreton et al., 2015; Lee, Daunizeau, et al., 2021). Computationally, decision-making has been conceptualised as a continuous process of evidence accumulation (Gold and Shadlen, 2001, 2007; Ratcliff and Rouder, 1998). Value-based choices have also been successfully characterised using evidence accumulation models (De Martino et al., 2013; Gluth et al., 2012; Krajbich et al., 2010; Pisauro et al., 2017; Polanı´a et al., 2015). Within this framework, the decision to gather more evidence before committing to a choice primarily depends on whether the accumulated evidence has reached a pre-specified decision threshold, implicitly linking expected accuracy and information seeking. From this perspective, individuals may learn fixed associations between confidence and behaviours, such as seeking information when confidence is low and refraining from doing so when confidence is high, facilitating generalisation across contexts. Although an advantage of this approach is that it offers a simple, generalisable rule (seek information when unsure), it can be misapplied, for instance because it does not take into account the value of further information, and thereby lead to suboptimal decisions.

Though the findings align with the tight link between metacognitive monitoring and control, with information seeking (a control process) primarily driven by confidence (a monitoring outcome) on a trial-by-trial level, they also suggest that control

processes might consider further variables, as evidenced by the flexibility in the relationship between confidence and information seeking at the level of condition (Experiment 1) or block (Experiment 2). Applying multivariate pattern analysis to fMRI data, different levels of metacognitive monitoring (measured as low to high confidence) were found to distinguish different acts of metacognitive control (measured as low to high desire for setting a reminder; Boldt and Gilbert, 2022). This overlap in brain patterns suggests a tight link between these two processes. However, the overlap was only partial, indicating that other signals, such as motivation or the costs and rewards associated with performance levels, may be incorporated into metacognitive control in addition to metacognitive monitoring signals.

The influence of various signals on cognitive control is also a key aspect of the Expected Value of Control (EVC) model (Shenhav et al., 2013). This model emphasises the flexibility with which control signals are allocated, based on the expected value of control in terms of the expected payoff, the required amount of control to achieve it, and the cost of exerting cognitive effort. The model suggests that this information is integrated in the dACC to determine the allocation of cognitive control. Shenhav et al. (2021) note, however, that there may be costs associated with engaging in the expected value of control calculation itself. This might lead people to rely on rough approximations or simplifying heuristics to determine when to exert control. Indeed, Bustamante et al. (2021) found that people learn the value of cognitive control based on stimulus features and rewards received, and then generalise the value of control learned in one situation to others with similar features, even when this is maladaptive. Consistent with this, participants in the current study may have broadly considered the costs and benefits of information seeking at a condition or block level, but settled on a default policy, whereby lower confidence generally signals a need for more information, at a trial level. My findings shed light on some of these intricacies in the EVC model. The evident and pronounced influence of con-

fidence on information seeking choices suggests that metacognitive control signals, such as confidence, are imperative inputs to the EVC model. The automated encoding of confidence during decision-making processes potentially acts as a primary input to determine the expected value of exerting control. Moreover, participants' behaviour in my experiments suggests that the EVC model might not always lead to optimal reward-maximizing behaviour. My observations, where participants seemed to adopt a general policy based on confidence without necessarily considering subtler decision-making factors, might suggest that such heuristic-driven behaviours could be a consequence of individuals trying to reduce the cognitive cost of continuously calculating the EVC in every single decision-making instance. My findings resonate with the idea that people may generalise the value of control based on past experiences, rewards, and stimulus features, suggesting that participants might have adopted a heuristic-based approach by generalising past experiences or context-related information on a broader scale (e.g., condition or block level), thereby potentially overlooking the unique subtleties of individual trials. The interplay between confidence and information seeking, the context-dependent flexibility in behaviours, and the potential inclination of individuals towards heuristic decision-making strategies all enrich the understanding of how the EVC model might manifest in real-world scenarios.

Shifting focus from the principal question about the flexibility in the relationship between confidence and information seeking, I also discovered effects on confidence itself. In both experiments, I observed a positive relationship between the absolute value of choice options and decision confidence, despite no observed increase in accuracy in decisions with higher value choice options. As there is no accuracy difference, greater confidence is not normatively justified. This increased confidence for choices involving higher value stimuli could be attributed to the possibility that value estimates leak into confidence judgements, in accordance with the idea that

the computations of confidence and value are inherently linked (Barron et al., 2015; De Martino et al., 2013; Lebreton et al., 2015). Research leveraging functional neuroimaging has reported neural correlates of confidence in the ventromedial prefrontal cortex (De Martino et al., 2013; Lebreton et al., 2015) as well as in mesolimbic and striatal regions (Guggenmos et al., 2016; Hebart et al., 2016). These regions comprise a brain network associated with the encoding of economic, motivational, and affective values (Levy and Glimcher, 2012). Such neural overlap between confidence and values implies a possible interaction at the behavioural level. According to this interpretation, a decision-maker's reported confidence could be higher not merely because they believe their decision is correct, but also due to the decision being made in a context of high expected or experienced value. This interaction between value and confidence could also underlie the observed associations between positive affective states and overconfidence (Allen et al., 2016; Giardini et al., 2008; Jönsson et al., 2005; Koellinger and Treffers, 2015; Massoni, 2014).

Another, not mutually exclusive, possibility is that when individuals report confidence, they may be expressing their expectations regarding the quality of the outcome, rather than merely the probability of their decision being correct. In this case, participants of the current experiments might have learned from previous trials that choices between options of higher value generally, on average, yielded greater winnings, thereby influencing their confidence judgements in subsequent similar trials. This interpretation aligns with findings showing that people integrate multiple past experiences to form global self-performance estimates (Rouault et al., 2019; Rouault and Fleming, 2020), and past confidence experiences can shape confidence expectations for future decisions (Boldt, Schiffer, et al., 2019). Regardless of the interpretation, my findings provide further evidence that incidental or contextual factors can influence confidence judgements beyond the immediate balance of evidence associated with the decision (Schwartz and Díaz, 2014).

A possible limitation of my study is the reliance on perceptual stimuli rather than real-world objects, which are frequently employed in value-based decision-making research (Brus et al., 2021; De Martino et al., 2013; Folke et al., 2016). That is, I used stimuli whose colour indicated their value, rather than the value being inherent in the stimulus. However, this approach allowed for careful control of the information upon which a decision is made. I was able to maintain consistent stimulus uncertainty by keeping the stimulus variance constant (i.e., uncertainty about each option's value individually) while only manipulating decision confidence through altering the choice option value difference between stimuli (i.e., uncertainty about an option's value relative to that of the other option; Pouget et al., 2016). Nevertheless, additional research that extends beyond perceptual decisions is warranted to investigate the degree to which my findings reflect the use of confidence in different settings.

To conclude, this study provides a more nuanced understanding of the interplay between confidence and information seeking. The findings establish converging evidence that underscores the context-dependency and strategic adaptability of this relationship, suggesting that decision makers modulate their information seeking behaviours according to the perceived value of additional information. This perceived value is not solely confined to the potential increase in accuracy derived from the information, but also extends to the prospect of increased reward resulting from such enhanced accuracy. Despite illustrating a certain degree of flexibility in information seeking behaviour, the study also uncovers a reliance on heuristic strategies and learned associations. Consequently, it suggests that individuals may not be engaging in sophisticated cost-benefit analyses for information seeking but instead might be employing simpler approximations to accommodate factors beyond confidence when determining the necessity for further information.

3

Confidence Matching in Social Settings

Chapter Abstract

When jointly performing a decision-making task, partners tend to report increasingly similar confidence judgements over time. However, the mechanisms underlying this phenomenon, termed '*confidence matching*', are not fully understood. Two primary potential drivers are postulated: informational factors causing individuals to genuinely recalibrate their confidence based on their partner's confidence; and normative factors leading individuals to adapt the way in which they overtly express their confidence, without changing their internal assessment of their own performance. To examine these influences, I conducted two experiments examining the effects of both informational and normative factors on internal and overtly reported confidence. The results demonstrate that both factors can lead to confidence matching. In a setting devoid of feedback, participants matched their confidence reports with their partner's and modified their information seeking behaviour — a proxy for internal confidence — accordingly, pointing towards the role of informational

factors. Conversely, in a scenario in which feedback was readily available but a joint decision-making rule was enforced, participants aligned their confidence reports with their partner's but did not adjust their information seeking behaviour, hinting at normative factors influencing the outward display of confidence matching. These findings highlight the flexibility and context-sensitivity of confidence, thereby underscoring the importance of factoring in social contexts and the adaptive nature of confidence when studying metacognitive processes.

3.1 Introduction

The ability to estimate our confidence in the decisions we make is a key feature of adaptive behaviour. People have been shown to reliably judge their own performance in the absence of feedback, with confidence reports closely tracking objective accuracy (Baranski and Petrusic, 1994; Koriat and Goldsmith, 1996). When communicating with others, we often share and discuss how confident we feel about our decisions or beliefs, and when making decisions together, self-reported confidence can be used to identify who within a group is most likely to know the correct answer (Pescetelli et al., 2016), such that sharing these self-evaluations can help to promote effective group behaviour (Shea et al., 2014).

Although an individual's reported confidence is typically proportional to the probability that their decision or belief is correct, various contextual factors may impact this relationship. For instance, factors such as mental health (Fu et al., 2012), physiological arousal (Allen et al., 2016; Jönsson et al., 2005), and current mood (Koellinger and Treffers, 2015) have been shown to impact confidence reports. Hence, the same absolute expression of confidence may have different meaning under varying circumstances. Moreover, variations in confidence expression extend beyond these transient influences. Confidence reports can further be shaped by individual-specific characteristics such as gender (Lundeberg et al., 1994) and personality (Campbell et al., 2004; Pallier et al., 2002; Schaefer et al., 2004). Ais et al. (2016) introduced the idea of a unique '*fingerprint*' in confidence profiles, asserting that the characteristics of confidence distributions are intrinsically tied to the individual, remaining remarkably consistent across different sessions and tasks. Furthermore, they observed that certain aspects of the confidence distribution, particularly an individual's confidence bias (mean confidence), can serve as an efficient indicator of a person's identity. This bias, interestingly, is also related to an indi-

vidual's optimism bias, bridging the gap between decisional and dispositional facets of confidence.

Given these various influences shaping an individual's expression of confidence, which are independent of the objective probability of being correct, joint decision-making under uncertainty can pose a challenging task. In theory, a robust approach to collective decision-making involves weighting individual opinions according to the confidence with which they are reported. Indeed, advice-taking studies have demonstrated that an advisor's confidence is the best predictor of their level of influence (Kuhn and Sniezek, 1996; Van Swol and Sniezek, 2002). However, if an individual's confidence is influenced by factors unrelated to their probability of being correct, these confidence variations, whether transient or stable, can disrupt the otherwise constructive impact of confidence-based opinion weighting. Even so, by exchanging confidence estimates with one another, joint decision-makers are able to outperform their respective individual performances and collectively improve sensitivity when making decisions (Bahrami et al., 2010). Importantly, this improvement hinges on the degree to which confidence is calibrated across group members (Silver et al., 2021) and there are also instances where communication may actually increase individual error rates (Kerr and Tindale, 2004; Valeriani et al., 2017).

Intriguingly, within the context of joint decision-making, there seems to be a propensity for partners to naturally align their confidence reports. That is, when jointly performing a decision-making task, partners match their confidence reports to each other, whereby over time, partners report more similar confidence judgements in terms of their average confidence as well as the distribution of confidence reports (Bang et al., 2017; Pescetelli and Yeung, 2020, 2022). This behaviour — termed '*confidence matching*' — is an effective strategy for maximizing joint accuracy when both partners have similar levels of expertise (Bang et al., 2017). Bang et al. (2017)

employed a signal detection model to simulate the dynamics of joint decision-making scenarios. Their model revealed that when simulated group members had equal levels of expertise, confidence matching enhanced joint accuracy. In further support of the model's findings, Bang et al. (2017) conducted an empirical experiment in which participants were paired with simulated partners of varying accuracy and confidence levels. In accordance with the modeling results, participants matched their confidence reports with their partners, and the extent to which joint accuracy improved depended on the initial group calibration. Although previous research has demonstrated the ubiquity of confidence matching, it is unclear whether it reflects a deeper change in one's internal computation of a sense of confidence or a surface change at the level of reporting confidence.

This idea of confidence potentially shifting either internally or merely at the overt reporting level in response to social influences finds its parallels in social psychology research on conformity. Two pivotal concepts are informational and normative social influences, stemming from seminal studies by Sherif (1935) and Asch (1951), respectively. Informational social influence refers to the psychological process whereby individuals mimic the actions of others, assuming these to reflect the correct behaviour in an uncertain situation. In such cases of uncertainty, individuals are inclined to adopt the group's beliefs or behaviours, viewing them as a reliable source of information. Conversely, normative social influence pertains to conformity, where individuals adjust their observable beliefs or behaviours to coincide with a group, even if these beliefs or behaviours contradict their private convictions. This influence is driven by a desire for social acceptance and fear of social rejection.

Expanding on these ideas, it is conceivable that confidence matching arises from individuals truly feeling more (or less) confident when interacting with partners who express high (or low) confidence. This view suggests that confidence matching hap-

pens at the internal level and is driven by informational conformity (Sherif, 1935), whereby a partner's confidence is used as a source of information to calibrate one's own confidence. One mechanism to consider is the Bayesian updating of '*priors*'. In a Bayesian framework, an individual starts with a prior belief about a situation. As they receive new data or evidence, they update this belief to form a more accurate posterior belief. When a partner expresses their confidence, this may influence an individual's prior about the task's complexity. A partner's confidence can serve as an informative gauge of how complex or challenging a task might be. If a partner expresses high confidence, it may signal that the task is less difficult than initially perceived, thus potentially elevating the individual's confidence. Conversely, if a partner expresses low confidence, it may suggest that the task is more challenging, leading to a potential decrease in the individual's confidence. According to Bahrami et al. (2012), sharing subjective confidence estimates could enhance social learning. In the absence of feedback, confidence can instead be utilised as a reinforcement signal. Although confidence cannot entirely replace objective feedback, it still provides a rough but useful estimate of the actual state of the world. By sharing confidence with others, access to more than one reinforcement signal is gained, ultimately resulting in collective benefits. Further to this, results from a study by Mahmoodi et al. (2019) suggest that people use others' confidence judgements to calibrate their own. When observing a worse performing partner, participants became more confident in their own performance, whereas conversely observing a partner with more accurate decisions lead to lower confidence. Similarly, Wittmann et al. (2016) showed that interacting with a better partner led to an overestimation of own abilities in a co-operative context and an underestimation of own abilities in a competitive context, whereas the opposite was true when interacting with a worse partner. Hence, it seems that indeed social information is integrated as a cue for calibration of one's sense of confidence.

Alternatively, it may be the case that people do not actually compute their likelihood of being correct differently when interacting with different partners, but only alter the way they express their confidence. This account implies that confidence matching occurs at the reporting level and stems from normative conformity (Asch, 1951). This implies that there are separable private and public expressions of confidence (Bang et al., 2020): the private representing one's genuine, internally computed confidence, and the public reflecting the overtly communicated confidence that is tailored to the social context. Take common social situations like job interviews or public presentations as examples. Even when someone might privately feel uncertain or anxious (private confidence), they often make an effort to project an image of confidence and certainty (public confidence) to appear competent and persuasive to their audience. In this case, it is not the computation of confidence itself that is altered. Instead, only the overtly expressed confidence is adjusted in response to the social context. This approach necessitates an additional cognitive process whereby an internal confidence variable is first computed and then mapped onto an overt confidence report — with the mapping depending on the social context. For instance, Mahmoodi et al. (2018) found that the degree to which participants were influenced by a partner's opinion was affected by how much the partner was influenced by their own opinion, serving as a signal of reciprocity. When making decisions with others, the decision expressed with higher confidence typically carries more weight (Zarnoth and Sniezek, 1997). Therefore, matching overtly expressed confidence may be driven by social norms such as reciprocity or an attempt to distribute responsibility evenly among group members. In line with this, it has been shown that people exhibit an equality bias, giving equal weight to each others' beliefs, regardless of actual differences in reliability (Mahmoodi et al., 2015). Some tasks that demonstrate confidence matching explicitly incorporate this incentive structure (Bang et al., 2017; Bang et al., 2020). They adopt the decision of the most confident individual as

the group's decision, providing individuals with a motive to match their confidence levels so that their opinions are not consistently overlooked or are not always dominating. Likewise, confidence matching could also result from mere imitation as it has been shown that people imitate and align each other's communication style in terms of syntax, semantics, and phonology (Fusaroli et al., 2012; Pickering and Garrod, 2004).

In drawing parallels between the Asch (1951) studies and confidence matching, it becomes evident that both phenomena involve a modification in people's overt expressions without changing their internal thoughts or beliefs. In the original Asch (1951) studies, participants publicly aligned with an incorrect group consensus despite privately recognising the truth, whereas in confidence matching, the public projection of certainty deviates from private sentiments. Moreover, a common driving factor in both scenarios is the presence of another person, emphasising the inherently social nature of the conformity observed. However, an essential distinction arises in the motivation behind these behaviours. In the Asch (1951) studies, normative conformity largely emerges from a fear of social rejection and a pressure to fit in with the majority. Conversely, in the context of confidence matching, the rationale behind adjusting one's overt confidence is not rooted in the fear of rejection, but rather in a desire to share responsibility with another person, ensuring mutual influence and collaboration in decision-making (El Zein et al., 2019). From hereon, when referring to '*normative conformity*' in this context, I specifically allude to this latter variant of '*normative*'.

These two accounts of confidence matching are depicted in Figure 3.1. When confidence matching is driven by informational factors, individuals adjust their confidence reports to align with their partner's confidence because their own internal confidence truly shifts in response to their partner's expressions of high or low confidence —

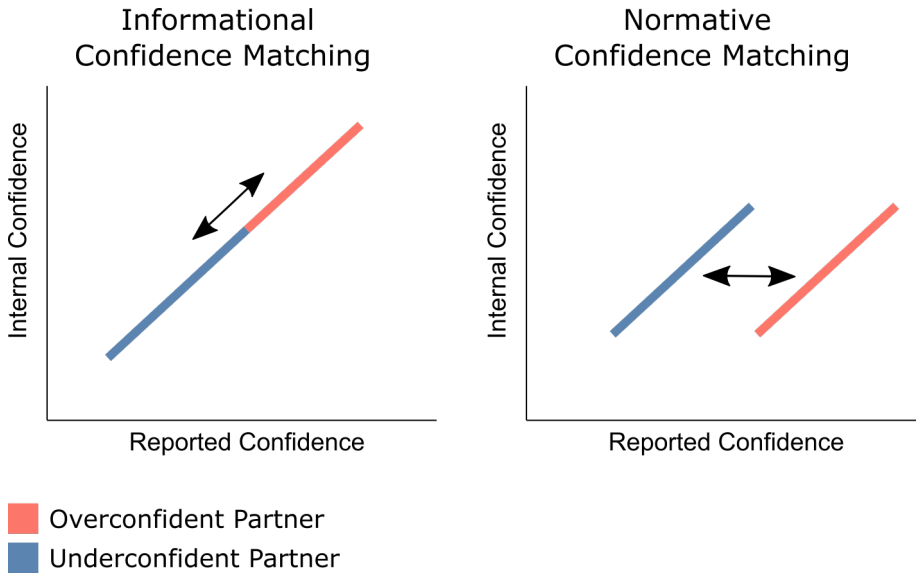


Figure 3.1: Confidence matching driven by informational versus normative factors. Under the informational confidence matching view, an increase in reported confidence from being paired with an underconfident partner to being paired with an overconfident partner would be in line with an increase in true internal confidence. Under the normative confidence matching view, this confidence increase would be at the report level only, causing a change in mapping between internal and overtly reported confidence.

that is, they feel more or less confident according to the social information received, and their reports reflect this change in internal confidence. As a consequence, the overall relationship between internal confidence and reported confidence is preserved — the points fall on the same line relating the two in Figure 3.1 — so that both are impacted proportionally by social information. Conversely, when confidence matching is driven by normative factors, participants' feelings of confidence do not change, but rather only the way they translate that internal confidence into a public report does. Thus, a change in the relationship between internal confidence and publicly reported confidence is necessary in response to interacting with different partners. As a consequence, for a given level of experienced conflict, a person's reports will vary according to whether they are interacting with a low versus high confidence partner.

Importantly, these explanations are not mutually exclusive. The experiments re-

ported in this chapter were motivated by the hypothesis that both explanations play a role in confidence matching, and that there are both informational and normative influences on internal and reported confidence. More specifically, I propose that informational factors cause confidence matching in the absence of feedback, when observing another person's confidence judgements can be used as a benchmark to gauge one's own performance. Conversely, normative factors come into play when social norms or joint decision rules, such as confidence-weighted aggregation of opinions, are introduced.

To test these hypotheses, participants in this study completed a perceptual decision-making task alongside virtual partners, under two conditions. In one condition, participants did not get any feedback but simply observed a partner's choices and confidence reports. In the other condition, participants received trial-by-trial feedback and their response and that of their partner were entered into a joint decision rule. Under this rule, the decision made with higher confidence was selected as the joint decision and in turn determined the reward earned by the participant on a given trial. A critical component of this experimental design was the inclusion of solo trials, on which instead of seeing their partner's response, participants had the option to seek additional information before finalising their decision. This design feature permitted me to use participants' information seeking choices as a behavioural proxy of their internal confidence. By hypothesis, this metric could potentially offer insight into a distinction between the participants' internal confidence and their overtly reported confidence. The use of information seeking behaviour as a measure of internal confidence is supported by previous research. Studies have found that neural correlates of confidence can predict information seeking choices (Desender, Boldt, et al., 2019), and that confidence reports can anticipate when individuals decide to revisit evidence (Desender et al., 2018) or seek advice (Pescetelli et al., 2021). Importantly, these findings hold true even when objective performance accuracy is

controlled for.

I anticipated that both conditions would lead to overt confidence matching, but with distinct implications for information seeking behaviour. This leads to two key predictions: (1) In the informational condition (but not the normative one), matching confidence with a partner's confidence reports should change participants' information seeking behaviour; (2) In the normative condition (but not the informational one), matching confidence with a partner's confidence reports should change the relationship between confidence reports and information seeking behaviour. These predictions are illustrated in Figure 3.1, which depicts the effects of the partner on overtly reported versus internal confidence. Here, information seeking is used as proxy for the '*internal confidence*' variable.

This chapter presents findings from two experiments. Experiment 2 was conducted not only with the objective of replicating the initial findings but also to mitigate order effects that were apparent in Experiment 1. Those participants of Experiment 1 who first experienced the condition with (assumed) informational influences, followed by the condition with (assumed) normative influences, exhibited the anticipated results. However, in contrast, those who were first subjected to the normative condition did not demonstrate an influence of the partner's responses on their information seeking behaviour within the subsequent informational condition. I hypothesised that this inconsistency may be attributed to the formation of a stable performance estimate for the perceptual decision task by the participants based on the feedback they received in the normative condition, thus rendering their confidence less prone to influence in the following informational condition. To counteract this, in Experiment 2, I introduced a second perceptual decision task, ensuring that participants were presented with varying tasks in each condition. As will be detailed in the results section, this modification successfully mitigated the order effect. Given

that the experiments are identical except for this modification of the first-order task, I present the results from both experiments together.

3.2 Methods

Participants. Participants were recruited through the online recruitment platform Prolific (www.prolific.co). Experiment 1 was completed by 30 participants (17 female, mean age ± SD: 31.3 ± 9.5). Experiment 2 was completed by 60 participants (41 female, mean age ± SD: 30.2 ± 8.2). Participants were compensated with a base amount of £10/hour plus an additional reward depending on performance in the tasks. The experiments were approved by the University of Oxford's Medical Sciences Interdivisional Research Ethics Committee.

Procedure. The experiments in this study consisted of two sessions that were completed by participants on two separate days, 2-7 days apart. The experimental design is depicted in Figure 3.2. Each session involved a series of trials of a perceptual decision task. In Experiment 1, participants performed the same dot discrimination task (depicted in Figure 3.2A) in both sessions. In Experiment 2, participants completed the dot discrimination task in one session and a colour discrimination task (depicted in Figure 3.2B) in the other session (order counterbalanced; see below for a description of these tasks). The two experiments were otherwise identical. On each trial, participants reported their decision and then rated how confident they felt in that decision. On a third of trials, participants then had the option to seek further information before committing to a final choice with no explicit cost. On all other trials, participants' choice and confidence rating were followed by the display of a partner's choice and confidence. Participants were in total paired with two partners in each session. Participants were told that the partners were past participants completing the same task, who had performed similarly well to them during the

practice trials. One of the two partners was designed to be '*underconfident*', whereas the other partner was designed to be '*overconfident*' (see details below). When making their initial response, participants did not know whether they were in an information seeking trial or in a partner trial.

In one of the two sessions of an experiment (order counterbalanced), participants completed a normative task version as depicted in Figure 3.2B. On information seeking trials, participants were rewarded based on the accuracy of their final decision. On partner trials, the reward was determined by the accuracy of the joint decision made with their partner. The joint decision was arbitrated in favour of the decision reported with higher confidence. Given this decision rule, the optimal strategy is to report confidence in a way that maximizes the probability of the joint decision being correct. Therefore, participants must adjust their confidence reports based on their partner's reports. For instance, if a participant is confident about the accuracy of their decision, they would need to report a relatively high level of confidence, such that their choice prevails in the joint decision. However, this dynamic becomes more complex if their partner generally reports high confidence due to their confidence bias. In this case, the participant would need to increase their confidence reports further beyond their usual levels to ensure that their decision is selected when they are fairly certain that they are indeed right. Conversely, if a participant is unsure, they should report lower confidence to allow their partner's choice to drive the joint decision. Yet, if the partner typically reports high confidence, even when unsure, the participant may not need to lower their confidence report even further than they normally would. Feedback was provided on every trial. In addition, after each set of 25 trials, participants received information about the average accuracy of their own decisions, their partner's decisions, and the joint decisions over the block.

In the other session of an experiment, participants completed an informational task

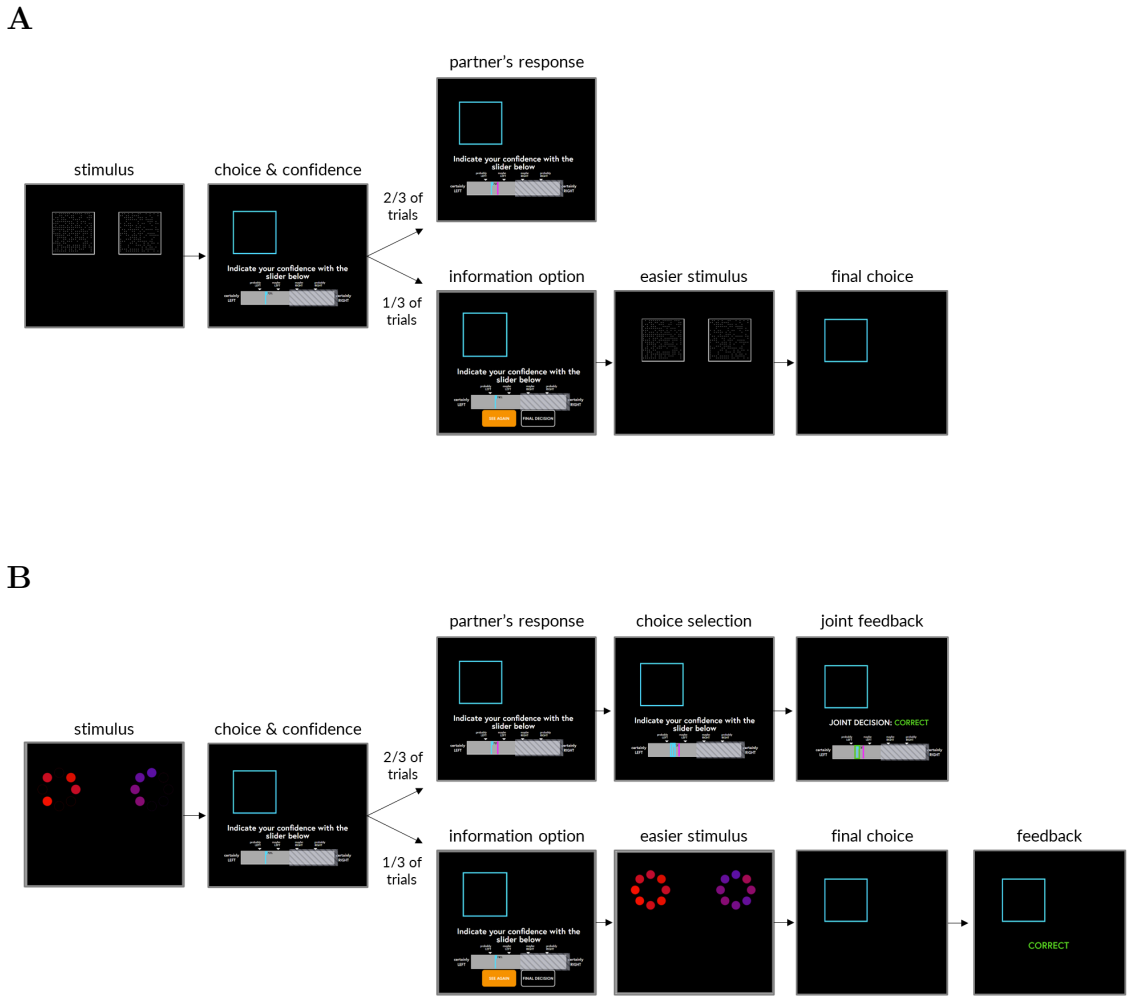


Figure 3.2: Experimental design. Participants performed either a dot discrimination task (Panel **A**) or a colour discrimination task (Panel **B**), with task assignment counterbalanced across participants. Following an initial decision, participants reported their confidence in the decision. Subsequently, a partner's response and confidence judgement was revealed in two-thirds of the trials. The unique feature of the normative task version (Panel **B**), as compared to the informational task version (Panel **A**), is that the choice reported with higher confidence was taken as the joint decision, with accompanying objective feedback. In one-third of trials, after the initial choice and confidence judgement, participants received a prompt to either seek additional information or progress to the next trial. If additional information was sought, a simplified version of the stimulus was presented, allowing for a final decision. Objective feedback was provided only in the normative task version (Panel **B**) and not in the informational task version (Panel **A**). In these examples, the dot discrimination task appears in the informational condition and the colour discrimination task is used in the normative condition, but for half of the participants this task-condition assignment was reversed.

version as depicted in Figure 3.2A. Just as in the normative task, participants observed their partner's decision and confidence after having reported their own. However, there was no joint decision stage and, importantly, no feedback. Instead, participants were rewarded based on the accuracy of their decisions as well as the precision of their confidence reports according to a Brier score (Brier, 1950). In the case of information seeking trials, only the final decision was considered for reward. The Brier score is calculated as the the squared difference between confidence rating c and decision accuracy o:

$$BS = \sum_{i=1}^{n}(c_i - o_i)^2 \quad \text{where} \quad o \in \{0, 1\}$$

Under this scoring rule, it is in participants' best interest to report their true belief about the probability of their decision being correct. Essentially, the Brier score is designed to penalise the differences between predicted probabilities (in this case, the confidence ratings) and the actual outcomes (the decision accuracy): the lower the Brier score, the better the predictions match the outcomes. Thus, participants are incentivised to reveal their true confidence to achieve a lower Brier score and maximise their rewards.

Tasks. In the dot discrimination task (performed for both sessions in Experiment 1, and for one session in Experiment 2), participants were asked to decide which of two boxes contains a larger number of dots and rate their confidence in their judgement. Each trial began with a fixation cross displayed for 1000 ms, followed by the presentation of two black boxes with an unequal number of white dots randomly positioned inside. After 300 ms, the dots disappeared and participants were prompted to make a choice using their mouse, by choosing the left box with a left click or the right box with a right click. After having made a decision, participants had to use their mouse to indicate their confidence on a visual analogue scale ranging from 50% (guessing) to 100% (completely certain that the chosen side has more dots). One

of the boxes (position counterbalanced) contained 200 dots and the other contained $200 + d$ dots, with $d \in [0, 100]$.

At the beginning of the experiment, participants completed 100 staircasing trials. During these trials, task difficulty was staircased to ca. 72% accuracy by titrating the d parameter. The staircasing procedure adjusted the step size on a logarithmic scale (i.e., larger adjustments for higher values of d) based on the log-adaptive two-down one-up procedure outlined in Rouault, Seow, et al. (2018). The starting step size was set at 4.6, which corresponded to a change of 100 dots. The step size was then modified by an amount of ± 0.4 for the first 5 trials, ± 0.2 for the next 5 trials, and ± 0.1 for the remainder of the staircasing trials. The step size was the same for both increases and decreases in the number of dots. When the task trials started, the d parameter was fixed to the average of the smallest d value (the hardest level) that the participant had ever achieved during the staircasing trials, and the final d value that was reached at the end of the staircasing trials.

If participants opted to see additional information on the information seeking trials, a fixation cross was again presented for 1000 ms followed by an easier version of the dot stimuli whereby the correct side remained the same but the dot difference d was increased by the number of dots equal to a step size of $+0.5$ in log-space.

In the colour discrimination task (performed for one session in Experiment 2), participants were asked to decide which of two stimuli was more red on average. Each stimulus was a multi-item display composed of eight coloured shapes spaced regularly in a circle, four of which were initially masked. The mean colour of all eight colour points varied in RGB space between pure red ($[1, 0, 0]$) and pure blue ($[0, 0, 1]$) and was determined by C with the mean colour corresponding to $[C, 0, 1 - C]$. The variance was fixed at 0.1. Similar to the dot discrimination task, each trial began with a fixation cross displayed for 1000 ms, followed by the presentation of two stim-

ulus displays. After 500 ms, the stimuli disappeared and participants were prompted to make a choice using their mouse by choosing the left stimulus with a left click or the right stimulus with a right click. After having made a decision, participants used their mouse to indicate their confidence on a visual analogue scale ranging from 50% (guessing) to 100% (completely certain that the chosen side is more red). Task difficulty was staircased to ca. 72% accuracy titrating the difference in C parameters between the two stimuli in the same fashion as for the dot discrimination task. If participants chose to seek information, they would again be presented with a fixation cross followed by stimuli in which all eight colour points for each stimulus were visible.

Computer-generated partners. Although participants were told that they were paired with past participants, the partners were in fact computer-generated. In each session, participants were paired with two different partners: an underconfident partner and an overconfident partner. Participants completed five blocks of 25 trials in one partner condition followed by five blocks in the other partner condition (order randomised). The first block with each partner served as a familiarisation block and did not include information seeking trials. On each trial, the partner's probability of choosing the correct response was determined as

$$p(correct) \sim \mathrm{N}(\mu = 0.8, \sigma = 0.066)$$

$$\text{where} \quad 0.6 \leq p(correct) \leq 1$$

such that partners were slightly more accurate than participants themselves in order to account for egocentric discounting, a bias whereby one's own opinion is weighted more than someone else's opinion (Yaniv and Kleinberger, 2000). Partners' confidence c was then computed as:

$$c = 50 + 50 \times (a + (p(correct) - 0.5) \times 0.8 + e)$$

$$\text{where} \quad \begin{cases} a = 0.1 & \text{for the low confidence partner} \\ a = 0.6 & \text{for the high confidence partner} \end{cases}$$

$$\text{and} \quad e \sim \mathrm{N}(\mu = 0, \sigma = 0.05)$$

By this design, the confidence of the partners was realistically calibrated, meaning that their reported confidence genuinely correlated with their objective accuracy. This approach is crucial, given that studies have shown people to be sensitive to such calibration (Pescetelli et al., 2021; Tenney et al., 2007)

In the equation, partners' confidence starts at a baseline of 50% (which signifies guessing) and can go up to 100% (representing complete certainty). This confidence is generally proportional to the objective probability of the decision being correct ($p(correct)$), but the relationship is imperfect. The scaling factor 0.8, adjusts how much the confidence c changes for a unit change in $p(correct)$. The offset term a modifies the confidence value in accordance with the overall tendency towards low or high confidence, differentiating between the underconfident and overconfident partners. Finally, e introduces further noise into the confidence reports, reflecting that confidence judgements in real life are not perfectly predictable or stable. As a result, the underconfident partner's reports were centered around 67% confidence, while the overconfident partner's reports centered around 92% confidence, thus creating distinct confidence profiles for each partner. Upon completing a set of five blocks with a particular partner, participants were asked to rate the partner's confidence, accuracy, and likeability on a ten-point Likert Scale. Additionally, participants were asked to indicate what gender they associated with their partner.

3.3 Results

Before the key analyses relating to confidence matching effects, I first report basic analyses of accuracy, confidence and information seeking choices, to establish that these were generally well-matched across task and partner conditions.

Accuracy and confidence calibration. Participants performed at the same accuracy level of 73% for initial choices in the normative and the informational task sessions ($t(29) = 0.3, p = .76$ for Experiment 1; $t(59) = 1.4, p = .16$ for Experiment 2). In both tasks, participants' confidence reports were well calibrated to their accuracy, with systematically higher confidence reports for correct compared to incorrect responses ($M = 75.42$ versus $M = 70.72$, $t(29) = 8.5, p < .001$ for the informational task, and $M = 77.62$ versus $M = 72.79$, $t(29) = 9.4, p < .001$ for the normative task in Experiment 1; $M = 76.58$ versus $M = 72.02$, $t(59) = 9.8, p < .001$ for the informational task and $M = 78.63$ versus $M = 73.04$, $t(59) = 10.2, p < .001$ for the normative task in Experiment 2). With first-order task performance being held constant, plotting Receiver Operating Curves gives an indication of participants' metacognitive sensitivity (the ability to distinguish between one's own correct and incorrect first-order responses). Participants displayed varying levels of metacognitive sensitivity, echoing findings on individual differences in metacognitive abilities (Fleming et al., 2010). Despite this variability, participants' metacognitive sensitivity was, on average, above chance (see Figure 3.3). This was confirmed by area under the ROC curve (AUROC) values significantly greater than 0.5 ($M = 0.66, t(29) = 12.6, p < .001$ when paired with the overconfident partner and $M = 0.65, t(29) = 16.4, p < .001$ when paired with the underconfident partner in Experiment 1; $M = 0.66, t(59) = 19.5, p < .001$ when paired with the overconfident partner and $M = 0.67, t(59) = 20.8, p < .001$ when paired with the underconfident partner in Experiment 2). There was no significant difference in

AUROC values between partner conditions ($t(29) = 0.3, p = .754$ in Experiment 1; $t(59) = 0.3, p = .781$ in Experiment 2)

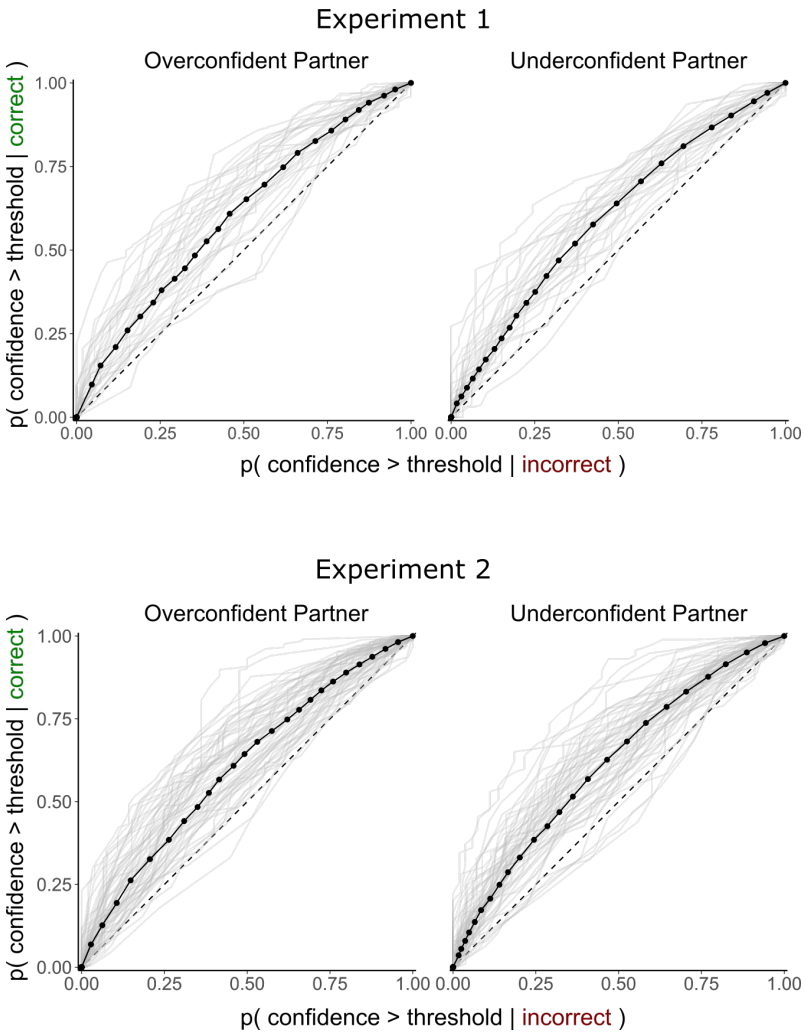


Figure 3.3: Metacognitive performance with each partner. Faint lines show individual participants' Receiver Operator Characteristic (ROC) curves. Solid lines show mean data across participants. ROC curves plot for each confidence value the probability of a confidence report being above that value for correct responses against the probability of a confidence report being above that value for incorrect responses. The area under the ROC curve (AUROC) is a measure of how well confidence ratings distinguish between correct and incorrect responses, whereby higher AUROC means higher metacognitive sensitivity with a value of 0.5 indicating performance at chance level (dotted line) and an AUROC of 1 indicating perfect discrimination.

Information seeking. On trials in which participants did not see their partner's

response but instead had the option to seek further information, participants chose to see the stimulus again in an easier version on 39.4% (Experiment 1) and 39.8% (Experiment 2) of trials. There was considerable variation between participants with information seeking ranging from 1% to 99% between participants for both Experiments, as is typically observed (Desender et al., 2018; Pescetelli et al., 2021). Participants made good use of further information with accuracy increasing from 68.1% for initial choices to 86.0% for second choices ($t(29) = 7.7, p < .001$) for Experiment 1 and from 67.4% for initial choices to 84.1% for second choices ($t(59) = 6.8, p < .001$ for Experiment 2) on trials in which participants chose to see more information.

Confidence matching. Despite accuracy being very similar across partner conditions ($t(29) = 1.5, p = .16$ for Experiment 1; $t(59) = 0.4, p = .71$ for Experiment 2), participants matched their confidence reports with the partners they interacted with, reporting overall higher confidence when paired with the overconfident partner and lower confidence when paired with the underconfident partner (see Figure 3.4). An ANOVA testing the effects of task condition (normative versus informational) and partner condition (overconfident versus underconfident) and their interaction showed a significant effect of partner condition on participants' reported confidence ($F_{partner}(1, 29) = 55.4, p < .001$, $F_{task}(1, 29) = 2.4, p = .13$, $F_{task \times partner}(1, 58) = 3.6, p = .07$ in Experiment 1; $F_{partner}(1, 59) = 69.2, p < .001$, $F_{task}(1, 59) = 3.9, p = .05$, $F_{task \times partner}(1, 118) = 2.7, p = .11$ in Experiment 2). Although the difference in reported confidence between partner conditions was numerically larger for the normative task, where participants have to match confidence to maximise reward, a substantial difference was apparent in both the normative task ($M_{shift} = 7.5$, $SE = 6.2$, $t(29) = 6.6, p < .001$ in Experiment 1; $M_{shift} = 7.6$, $SE = 8.3$, $t(59) = 7.1, p < .001$ in Experiment 2) and the informational task ($M_{shift} = 4.6$, $SE = 5.8$, $t(29) = 4.3, p < .001$ in Experiment 1; $M_{shift} = 5.7$,

$SE = 7.0$, $t(59) = 6.3, p < .001$ in Experiment 2), with substantial confidence shifts even exceeding those observed between correct and incorrect responses ($M_{shift} = 4.7$, $SE = 2.4$ in Experiment 1; $M_{shift} = 5.2$, $SE = 3.3$ in Experiment 2).

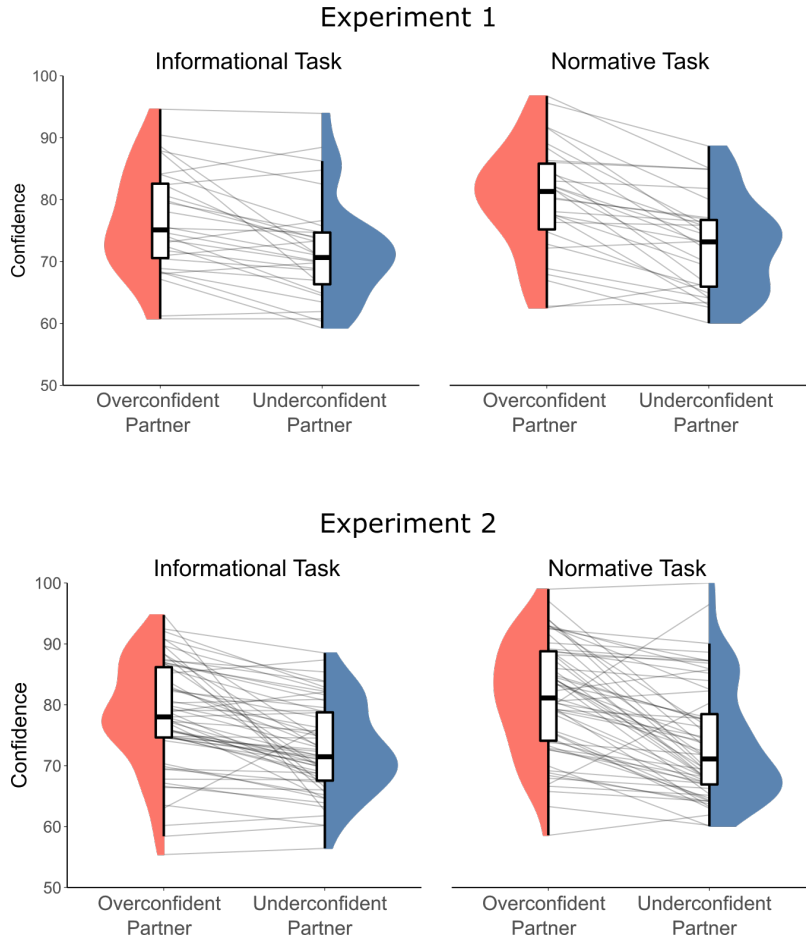


Figure 3.4: Confidence in the informational and normative tasks when paired with the underconfident and overconfident partner. Faint lines show individual participant means. Violin and box plots show the distributions of participant means.

Partner × task interaction effects. Given that confidence matching was successfully induced in both experimental contexts — that is, both with (presumed) informational and normative influence — the critical analyses concern whether the corresponding effects observed on participants' expressed confidence nevertheless mask crucial underlying differences in their internal confidence. As a first test of this

key hypothesis, that informational and normative influence should have differential effects on internal confidence, I examined how overconfident versus underconfident partners affected participants' information seeking choices across task conditions (normative versus informational). The results, shown in Figure 3.5, were analysed in an ANOVA testing the influence of task condition and partner condition and their interaction on the proportion of trials where participants decided to seek further information. My key prediction was that there should be a significant interaction effect between task condition and partner condition: if informational influence, but not normative influence, changes internal confidence, then we should see a change in information seeking behaviour between partner conditions in the informational condition only.

Indeed, this interaction was significant in Experiment 1 ($F_{task}(1, 29) = 2.5, p = .12$, $F_{partner}(1, 29) = 0.05, p = .82$, $F_{task \times partner}(1, 58) = 5.9, p < .05$), whereby participants tended to seek more information when paired with the underconfident partner compared to the overconfident partner in the informational task ($t(29) = -2.0, p = .06$) with the opposite pattern for the normative task $t(29) = 1.3, p = .21$). In Experiment 2, a numeric trend in the same direction was apparent, but not statistically significant ($F_{task}(1, 59) = 0.3, p = .56$, $F_{partner}(1, 59) = 0.4, p = .53$, $F_{task \times partner}(1, 118) = 2.8, p = .10$). Thus, the change in information seeking behaviour depending on the partner's confidence bias was evident as hypothesised for the informational task. Interestingly, information seeking behaviour in the normative task did not merely remain unaffected, but unexpectedly exhibited an inverse pattern compared to the informational task.

My second key analysis focused on the relationship between reported confidence and information seeking on a trial-by-trial level, evaluating whether the predicted relationship (and dissociation) between these two measures across partner conditions

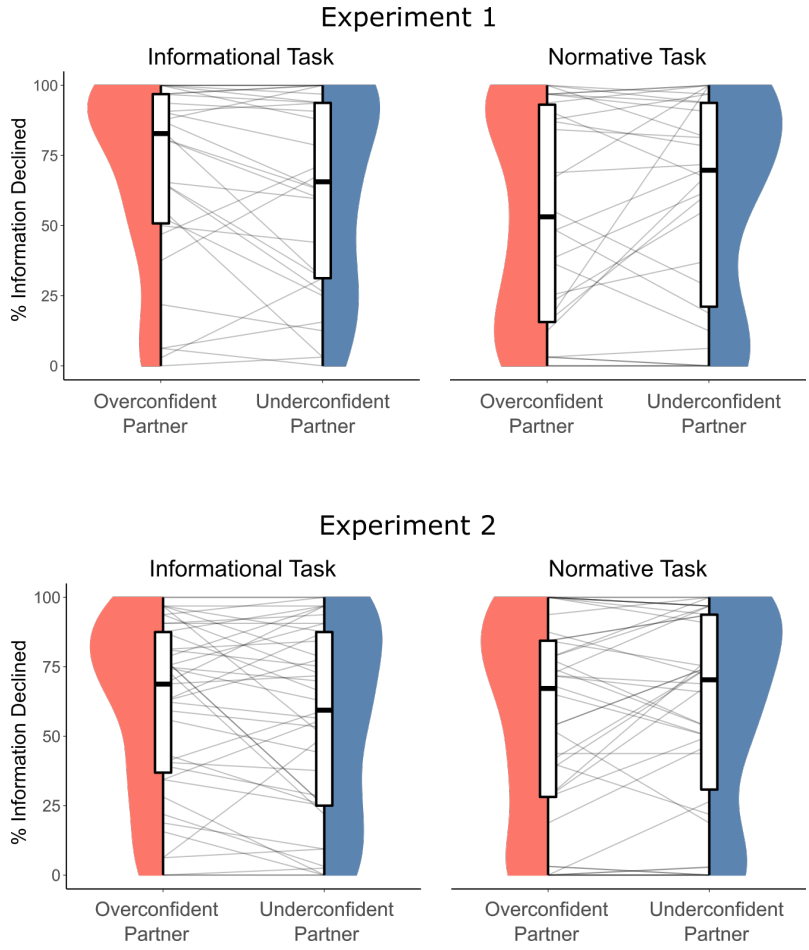


Figure 3.5: Information seeking in the informational and normative tasks when paired with the underconfident and overconfident partner. Faint lines show individual participant means. Violin and box plots show the distributions of participant means.

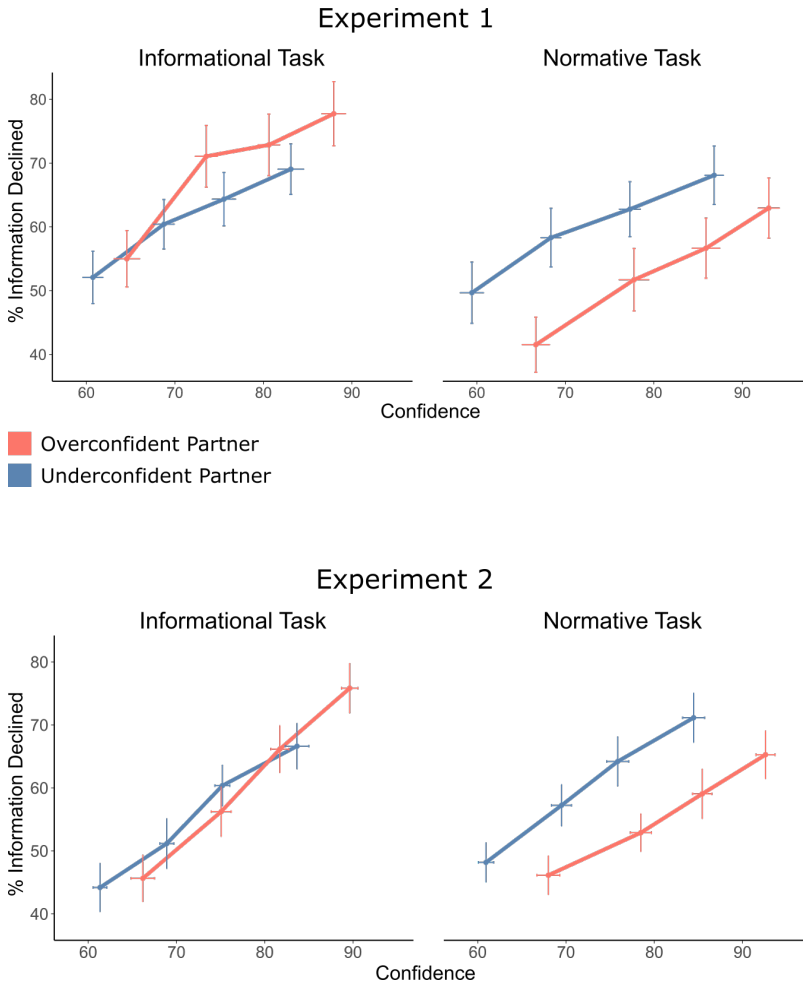


Figure 3.6: Information seeking as a function of reported confidence in the informational and normative tasks when paired with the underconfident and overconfident partner. For each participant, confidence was divided into quartiles and the proportion of information seeking choices within each confidence quartile was computed. The graph displays the grand mean across participants and standard error of the mean for both reported confidence and information seeking.

as depicted in Figure 3.1 would be observed. Aggregated results are shown in Figure 3.6, where y-axis values plot participants' likelihood of declining further information as the critical proxy for internal confidence. The empirical results overall match well with the patterns predicted in Figure 3.1. The analysis was performed on a trial level by applying a linear mixed-effects model (LMM). Trial-wise information seeking choices were predicted in a logistic regression by reported confidence, task condition, partner condition, and an interaction term between task and partner conditions, allowing for random intercepts and slopes by participant. There was a significant effect of confidence ($\chi^2 = 40.12, p < .001$ in Experiment 1; $\chi^2 = 169.87, p < .001$ in Experiment 2) as well as the interaction between task condition and partner condition ($\chi^2 = 9.02, p < .01$ in Experiment 1; $\chi^2 = 4.28, p < .05$ in Experiment 2) on information seeking. To follow-up on the interaction, I next ran separate regression models per task condition, predicting information seeking choices in a logistic regression by reported confidence and partner condition, allowing for random intercepts and slopes by participant. These analyses revealed that, when controlling for confidence, there was no effect of partner condition on information seeking in the informational task condition ($\chi^2 = 0.83, p = .36$ in Experiment 1; $\chi^2 = 0.68, p = .41$ in Experiment 2), whereas there was a significant effect of partner condition on information seeking in the normative task condition ($\chi^2 = 5.06, p < .05$ in Experiment 1; $\chi^2 = 10.22, p < .01$ in Experiment 2). These results are in line with my predictions. They demonstrate that while reported confidence drives information seeking behaviour across both task conditions, there is an additional impact of the partner condition in the normative condition. When paired with an underconfident partner, less information is sought than would be expected based on the level of reported confidence (i.e., suggesting that the internal confidence is in fact higher than the overtly reported value). Conversely, when paired with an overconfident partner, more information is sought than would be expected based on the level of

reported confidence (i.e., suggesting that the internal confidence is in fact lower than the overtly reported value).

To summarise, the findings align with key predictions, indicating the presence of distinct informational and normative impacts on internal and reported confidence. The ANOVA results highlight distinct partner influences on information seeking behaviour between the informational and normative task conditions, though this effect was only trending in Experiment 2. The regression analysis distinctly indicates a divergence between internal and reported confidence under normative social influences but not informational ones. In the informational task, accounting for reported confidence negates any effect of partner on information seeking behaviour, as the change in information seeking between partner conditions aligns with the participant's reported confidence. In contrast, in the normative task, for a given level of reported confidence, participants decline less information when paired with the underconfident partner compared to being paired with the overconfident partner. This pattern indicates that a higher level of internal confidence corresponds to a given level of reported confidence when participants interact with an underconfident partner compared to with an overconfident partner (see Figure 3.6).

Order effects. As described in the introduction, a motivation for conducting Experiment 2 was the observation of order effects in Experiment 1, whereby participants who first experienced the informational task condition followed by the normative task condition did show the anticipated effects, whereas participants with the opposite order of task conditions did not. To test for order effects, I ran the ANOVA testing the first key prediction now including task order alongside task condition and partner condition and all interaction terms as predictors of participants' information seeking. In Experiment 1, there was a significant interaction between task condition and partner condition ($F(1, 58) = 6.94, p < .05$), reiterating the finding

reported above showing that the influence of partner on information seeking differed between the informational and normative conditions, but there was also a significant three-way interaction between task order, task condition, and partner condition ($F(1, 87) = 6.28, p < .05$). No other effects were significant (all $p > .12$). As can be seen in Figure 3.7, for participants experiencing the informational task condition first, the results in this condition are as expected, with participants reporting higher confidence and declining more information when paired with the overconfident compared to the underconfident partner. However, for participants who started with the normative task condition, the adjustment in confidence reporting and information seeking behaviour no longer manifested in the informational task condition. This observation is evidenced by the substantial overlap between the blue (underconfident partner trials) and pink (overconfident partner trials) lines. One potential explanation for this observation is that participants might have developed a performance estimate for the dot discrimination task during the normative task condition, such that the partner's confidence no longer carries relevant information about task difficulty, and thus has no effect on participants' confidence and thereby their information seeking behaviour. In Experiment 2, as reported above there was a numeric trend for the interaction between task condition and partner condition, which however as not significant ($F(1, 118) = 2.74, p = .103$), and there was no significant three-way interaction between task order, task condition, and partner condition ($F(1, 177) = 0.15, p = .699$). No other effects were significant (all $p > .25$). As can be seen in Figure 3.7, for Experiment 2, the effects of partner condition on reported confidence and information seeking behaviour between task conditions are similar across the two condition orders.

Partner perception. After a sequence of five blocks interacting with a specific partner, participants evaluated the confidence, accuracy, and likeability of their partner using a ten-point Likert Scale. Additionally, they also identified

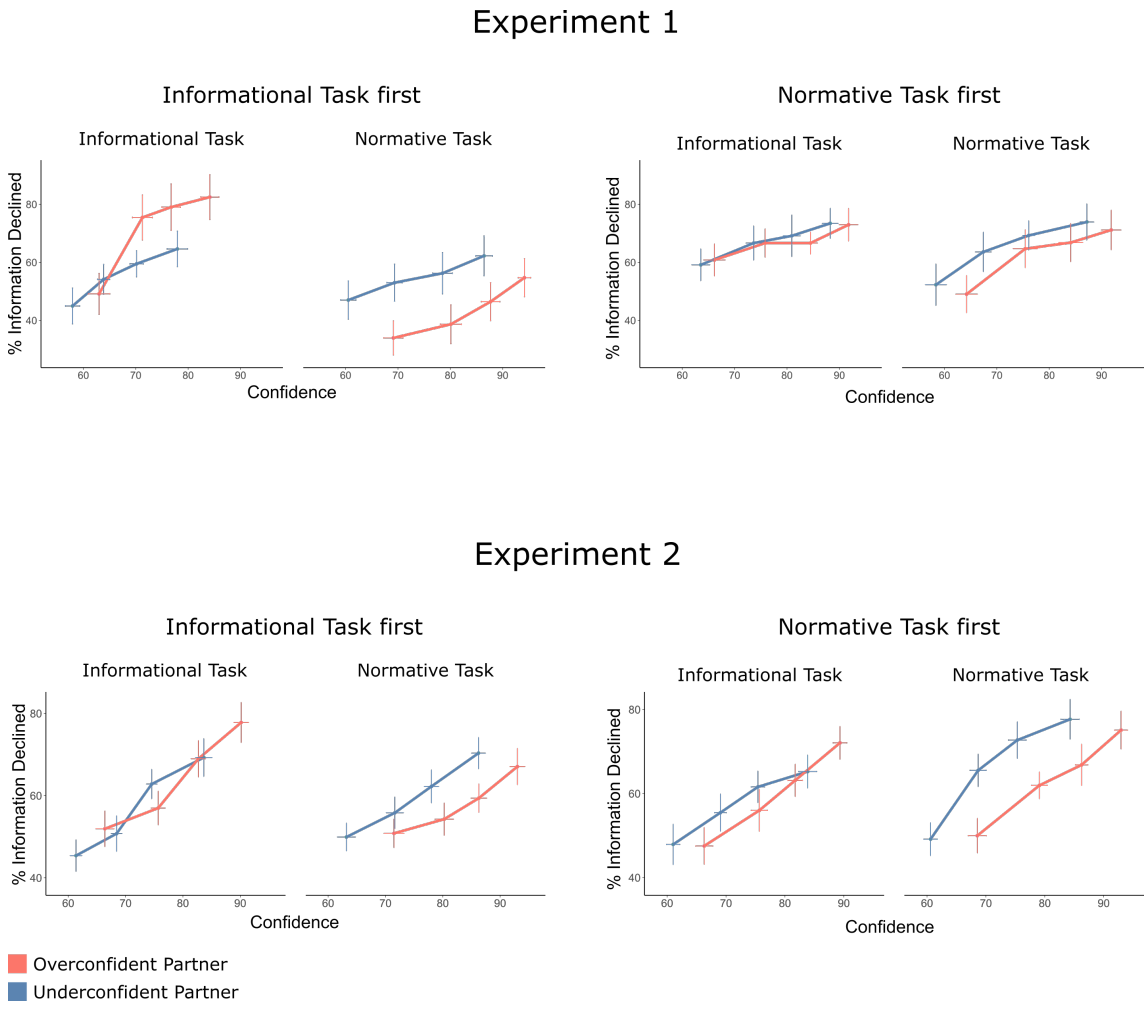


Figure 3.7: Order effects. Information seeking as a function of reported confidence in the informational and normative tasks when paired with the underconfident and overconfident partner for the two task orders. For each participant, confidence was divided into quartiles and the proportion of information seeking choices within each confidence quartile was computed. The graph displays the grand mean across participants and standard error of the mean for both reported confidence and information seeking.

the gender they associated with their partner. The mean ratings for these responses are summarised in Table 3.1. Across both experiments, participants perceived the overconfident partner as more confident than the underconfident partner. In Experiment 1, this effect was especially prominent in the informational task condition ($F_{task}(1, 29) = 0.003, p = .95$, $F_{partner}(1, 29) = 81.66, p < .001$, $F_{task \times partner}(1, 29) = 9.33, p < .01$ in Experiment 1; $F_{task}(1, 59) = 3.77, p = .06$, $F_{partner}(1, 59) = 115.78, p < .001$, $F_{task \times partner}(1, 59) = 0.24, p = .63$ in Experiment 2).

The partners were perceived as more accurate in the normative task condition compared to the informational one. Despite receiving objective feedback in the normative task, participants appeared to rely on the confidence-heuristic (Thomas and McFadyen, 1995), leading to a perception of the overconfident partner as more accurate than the underconfident one ($F_{task}(1, 29) = 11.40, p < .01$, $F_{partner}(1, 29) = 5.93, p < .05$, $F_{task \times partner}(1, 29) = 0.01, p < .95$ in Experiment 1; $F_{task}(1, 59) = 10.61, p < .01$, $F_{partner}(1, 59) = 3.73, p = .06$, $F_{task \times partner}(1, 59) = 0.84, p = .37$ in Experiment 2).

Likeability ratings did not significantly differ between the overconfident and underconfident partners. However, the task condition effect suggested that participants found their partners more likeable in the normative task ($F_{task}(1, 29) = 9.67, p = .01$, $F_{partner}(1, 29) = 1.98, p = .17$, $F_{task \times partner}(1, 29) = 0.12, p = .73$ in Experiment 1; $F_{task}(1, 59) = 7.99, p < .01$, $F_{partner}(1, 59) = 0.00, p = .95$, $F_{task \times partner}(1, 59) = 0.10, p = .76$ in Experiment 2).

Finally, participants were more likely to identify the overconfident partner as male compared to the underconfident partner ($F_{task}(1, 29) = 0.00, p = .99$, $F_{partner}(1, 29) = 7.27, p < .05$, $F_{task \times partner}(1, 29) = 1.68, p = .21$ in Experiment 1; $F_{task}(1, 59) = 2.19, p = .15$, $F_{partner}(1, 59) = 7.91, p < .01$, $F_{task \times partner}(1, 59) = 1.40, p = .24$ in Experiment 2).

A related study by Bang et al. (2017) involving a joint decision-making paradigm showed similar findings, whereby participants held a stereotype that highly confident partners were more likely to be identified as male and less confident partners as female. Furthermore, in their study, participants liked partners who were highly accurate yet displayed low confidence the most. However, in contrast, I found no significant influence of partners' confidence on their likeability.

Table 3.1: Mean (± standard deviation) ratings on a 10-point Likert Scale of perceived confidence, likeability, accuracy, and gender in Experiments 1 and 2 in the normative and informational tasks for underconfident (Underc. P.) and overconfident (Overc. P.) partners.

		Normative Task		Informational Task	
		Underc. P.	Overc. P.	Underc. P.	Overc. P.
Confidence	1	5.6 ± 2.0	7.8 ± 1.8	4.8 ± 1.8	8.3 ± 1.7
	2	5.7 ± 1.8	8.1 ± 1.8	5.0 ± 1.9	7.9 ± 1.8
Likeability	1	6.8 ± 1.8	6.6 ± 1.6	6.2 ± 1.8	5.7 ± 1.6
	2	6.9 ± 1.7	6.8 ± 1.8	6.1 ± 1.6	6.2 ± 1.4
Accuracy	1	6.8 ± 1.6	7.3 ± 1.7	5.7 ± 1.6	5.9 ± 1.8
	2	6.8 ± 1.7	7.1 ± 1.9	5.9 ± 1.6	6.1 ± 1.9
% male	1	46.4	64.3	35.7	76.9
	2	50.9	69.0	42.9	64.3

3.4 Discussion

The ability to evaluate one's confidence in decisions and beliefs is a crucial aspect of adaptive behaviour. Sharing these evaluations with others can enhance group decision-making and improve collective performance even in the absence of feedback (Bahrami et al., 2010; Shea et al., 2014). As partners work together, their average reported confidence as well as the distribution of confidence reports become increasingly alike (Bang et al., 2017; Pescetelli and Yeung, 2020, 2022). This confidence matching may be driven by an attempt to better calibrate one's own confidence judgements to the difficulty of the task, a form of informational influence (Sherif, 1935), affecting confidence at the internal level. Alternatively, confidence matching may be driven by social reasons unrelated to the accuracy of confidence judgements. Such factors may include a desire to fit in socially, to avoid being perceived as incompetent or, critically, to promote equality in group decision-making. These are all forms of normative influence (Asch, 1951), affecting confidence at the reporting level only. In two experiments, participants performed a series of perceptual decisions alongside a partner. The partner's confidence reports as well as the presence of feedback and need for joint decision-making were manipulated to better understand the conditions under which social information causes confidence matching at the internal or reporting level.

When provided with objective feedback and required to make joint decisions (normative task), participants strategically matched their confidence reports to those of their partners to maximise reward. However, their information seeking behaviour did not change when being paired with different partners. In the absence of objective feedback and the requirement for joint decision-making (informational task), participants also matched their confidence reports to those of their partners. In contrast to the normative task, participants did adapt their information seeking behaviour

when paired with different partners, tending to seek less information in the context of a high confidence partner.

Thus, although the same overt metacognitive behaviour, a matching of reported confidence, occurs in both cases (replicating results seen previously by Bang et al. (2017) and Pescetelli and Yeung (2020, 2022)), differences in partner effects on information seeking suggest that the nature of this confidence matching differs. Taking information seeking as a behavioural marker of internal confidence (Desender, Boldt, et al., 2019; Desender et al., 2018; Pescetelli et al., 2021), the absence of a shift in information seeking between partners in the normative task indicates that people strategically change the mapping from their internal confidence to overt confidence reports depending on the social context (Bang et al., 2017; Bang et al., 2020). By the same token, the presence of a shift in information seeking between partners in the informational task suggests that participants use their partner's confidence reports to inform and calibrate their own internal confidence. These findings highlight the context-sensitive influence of both normative and informational factors in shaping confidence reports as different mechanisms behind confidence matching in social contexts.

My results support the notion of confidence as a general abstract representation of an individual's likelihood of being correct, allowing for flexible and generalised metacognitive control over thought and behaviour (Yeung and Summerfield, 2012). Notably, in the normative task condition, participants exhibited this flexibility by using their confidence estimates to guide information seeking choices independent of how they mapped their confidence onto overt reports based on the social context. Although these results are as predicted, it is noteworthy that they imply that participants are able to clearly separate their representations of confidence as they relate to different actions, i.e., confidence as used to drive information seeking and confidence as used

to generate a socially suitable report. This result could not be taken for granted, even given my predictions, given that other studies suggest confidence '*leaks*' across decisions such that confidence expressed in one task influences confidence expressed in a subsequent, different task (Rahnev et al., 2015). The term '*leak*' describes the idea that confidence from one task is seeping into the next, rather than being contained within the specific task it originated from. The underlying assumption is that the human brain tends to expect some stability in the environment and therefore uses information from past experiences to interpret current situations. That is, the brain forms expectations about the reliability of current decisions based on past ones. However, despite this hypothesis of confidence leak, participants were in fact numerically more (less) likely to decline information, indicating higher internal confidence, when strategically reporting low (high) confidence, which is the opposite direction of effect to that expected if changes in reported confidence reflected changes in internal confidence.

This separation between an initial confidence computation and a subsequent context-dependent confidence report aligns with the findings of a recent fMRI study by Bang et al. (2020). This study found that the dorsal anterior cingulate cortex (dACC) and perigenual anterior cingulate cortex (pgACC) tracked internal confidence, whereas a context-dependent mapping to overtly reported confidence was associated with activity in dACC and lateral frontal pole (FPl). Moreover, a study by Boldt and Gilbert (2020) found partial but not complete overlap between patterns of brain activity associated with metacognitive monitoring versus metacognitive control. This separation may reflect the need to take into account additional factors, such as the social context, when exerting metacognitive control, over and above an initial confidence computation through a metacognitive monitoring process. These findings highlight the importance of taking into consideration the fact that confidence is used flexibly and adapting to the task at hand when studying metacognitive pro-

cesses, and that the way in which confidence is probed may impact the judgement reported.

My results suggest that models of confidence should allow for flexible adjustments of confidence according to contextual factors. As the presence of different partners affected confidence at the internal level in the informational task, contextual factors appear to be integrated into this signal rather than confidence merely being a read-out of the current state of evidence accumulated in the first-order decision (Boldt et al., 2017; Kiani et al., 2014; Koriat and Levy-Sadot, 2001; Murphy et al., 2015). Taking into account a variety of contextual information that is not directly related to the task at hand, but is known to co-vary with problem-specific evidence, can sometimes lead to mistakes and biases. However, it also enables quick decision-making even when information is scarce (Gigerenzer and Todd, 1999). This notion aligns with the Bayesian inference model for meta-memory proposed by Hu et al. (2021) which provides a normative framework for the integration of such contextual cues in confidence judgements. This model posits that individuals estimate their confidence based on both their processing experiences during a task and their pre-existing beliefs about their overall capabilities, suggesting that the incorporation of such cues should be seen as an adaptive feature rather than a '*bug*' in the system. Matching confidence judgements with another person is a simple and computationally efficient heuristic to help solve the inference problem '*how well am I performing at the task*'. This is true in particular when partners have similar levels of expertise, which is often the case as we tend to associate with people who share similar traits (McPherson et al., 2001).

A study by Rouault et al. (2019) proposes that confidence encoding occurs at multiple hierarchical levels with recent findings suggesting that individuals integrate numerous local, trial-by-trial confidence estimates to form global self-performance

estimates, which guide higher-order behavioural control (Boldt, Blundell, et al., 2019; Cavalan et al., 2023; Rouault et al., 2019). My findings suggest a reciprocal relationship between higher-order beliefs about self-performance and local confidence estimates. This relationship implies that higher-order self-performance beliefs impact local confidence estimates beyond the bottom-up evidence obtained on individual trials. From this perspective, observing a partner's confidence judgements may affect an internal prior on expected task performance, such that socially induced changes in local trial confidence are the result of a change in a global self-performance estimate. Work by Van Marcke et al. (2022) underscores the pivotal role of prior beliefs in shaping decision confidence. In their research, they revealed that manipulating prior beliefs about task performance in a training phase can causally shift reported confidence levels during a test phase, while keeping objective performance unchanged. Their findings underscore that confidence levels can diverge substantially due to varying prior beliefs about the ability to execute a task. This aligns with the notion that a change in prior due to witnessing a partner's confidence for a task may markedly shape ensuing confidence reports.

Although the current study provides compelling evidence highlighting the flexible use of social information in forming confidence judgements based on the context, the broader applicability of these findings needs further exploration. A study conducted by Wittmann et al. (2016) observed a similar pattern, showing how interaction with a superior partner resulted in overestimating personal capabilities in a cooperative setting, and underestimating them in a competitive one, supporting the notion of flexibility in the way in which social information may influence confidence judgements depending on the context. Further research might examine other dimensions of this flexibility. One potential area of exploration is whether the context-based effects akin to those outlined by Wittmann et al. (2016) represent genuine changes in confidence, or are instead reflective of communicative adjustments. Furthermore,

my research prompts additional inquiries into previous findings on how confidence judgements are affected by situational factors and individual differences, such as current mood (Koellinger and Treffers, 2015), gender (Lundeberg et al., 1994), or personality traits (Campbell et al., 2004; Pallier et al., 2002; Schaefer et al., 2004). A key question that emerges is the degree to which these results represent actual changes/differences in the internal sense of confidence, versus mere alterations in the overt communication of confidence.

In this study, I examined a rather focused and controlled social setting, which does not fully represent the complexities and nuances of genuine social interactions. For example, decisions made by participants in this setting did not reciprocally influence their partners, a departure from many real-life contexts where one's choices reverberate back and forth. If one were to extrapolate this to a live, collaborative environment with multiple participants making simultaneous decisions, the dynamics would likely become more complex. Factors such as instantaneous feedback, shifts in decision-making strategies, direct verbal and non-verbal social cues, and heightened social pressures could amplify the impact of one participant's confidence on others. Additionally, considering that decisions in the real world often involve more than just pairs, larger group dynamics could introduce even more layered mechanisms influencing the confidence levels and reports of participants. Although this study offers foundational insights into how social settings influence confidence assessments, broadening the scope to encompass diverse social dimensions, like competition, cooperation, or hierarchical dynamics, could shed more light on the multifaceted ways in which the social context shapes a decision-maker's confidence judgements.

In conclusion, the present study demonstrates that confidence matching in social settings can be driven by both normative and informational factors. Confidence emerges as a flexible, context-dependent tool used to optimise behaviour. The find-

ings highlight the importance of considering the influence of social context and the flexible nature of confidence when studying metacognitive processes. The presented methods demonstrate how these influences can be experimentally induced, paving the way for future research to address how these findings may be applied to improve group decision-making in real-world settings.

4

Neural Correlates of Confidence Matching

Chapter Abstract

This study builds upon the experiments reported in the previous chapter exploring the effects of informational and normative conformity on confidence matching in a social context. The primary objective was to deepen the understanding of potential contrasting effects of informational and normative social influences on internal confidence and overtly reported confidence judgements. To this end, I sought to identify a continuous neural marker of internal confidence via electroencephalography (EEG), marking a departure from the previous chapter in which information seeking behaviour was used as a proxy for internal confidence. The study design comprised a perceptual decision task in an informational task condition, where participants observed a partner's decisions and confidence ratings without objective feedback, and a normative task condition, where a joint decision was determined based on the highest confidence rating and feedback was provided. The behavioural outcomes of the study were largely consistent with the previous experiments, revealing confidence

matching in both tasks. However, the deployment of multivariate decoding techniques to establish a trial-by-trial internal confidence index from EEG data met with limited success. An alternative univariate analysis identified a significant correlation between the P3 component at the Pz electrode and confidence when experimental conditions were considered as covariates. However, no significant effects of task condition (informational/normative), partner condition (overconfident/underconfident partner), or their interaction on the P3 component emerged. Given the constraints in identifying a robust neural index of internal confidence, the findings remain inconclusive with regards to the role of informational and normative social influences on internal and reported confidence.

4.1 Introduction

The previous chapter described two experiments exploring confidence matching in social contexts. Specifically, these experiments showed that people aligned their confidence judgements with those of others in ways that reflected two distinct types of conformity: informational conformity (Sherif, 1935), impacting confidence at the internal level, and normative conformity (Asch, 1951), affecting confidence at the reporting level only. Using information seeking behaviour as an indirect measure of internal confidence, the previous experiments showed that, in scenarios in which feedback was absent, the exposure to others' confidence judgements affected not just the confidence reports overtly expressed by participants, but also the decisions they made regarding the amount of information they sought privately. In contrast, the introduction of feedback and a joint decision rule led to a divergence in the effects observed on overtly reported confidence and privately sought information. Under these conditions, participants only changed the reported confidence to match with the partners they interacted with, whereas their information seeking behaviour was unaffected.

In the process of investigating the impact of social information on confidence, the use of information seeking behaviour was useful as a proxy for participants' internal sense of confidence. The theory, whereby normative and informational social influences have different impacts on overtly reported and internal confidence judgements, also makes predictions about neural activity related to confidence. Therefore, this chapter aims to integrate and build upon the initial findings by adopting a neuroimaging-based approach. It specifically explores contrasting effects of different social contexts on overtly reported confidence and on single-trial electroencephalography (EEG) measures of neural activity related to confidence.

A limitation of the previous experiments' design — in which participants were given the option to view a less noisy version of the stimulus when unsure about their choice — is its binary nature: it does not tell us about the degree to which participants felt a need for additional information, but only whether this need passed a certain threshold. In other words, although information seeking behaviour can tell us if an individual feels uncertain, it does not provide a nuanced picture of the extent of their uncertainty — that is, their internal sense of confidence. Unlike the binary nature of information seeking, neural markers of confidence offer a continuous measure, which may capture the nuanced levels of internal confidence more accurately. Additionally, the previous study design imposed certain restrictions, limiting the opportunity to probe information seeking choices to just one-third of the trials. In contrast, EEG markers of confidence can be measured on every trial, thus avoiding the limitations imposed by the design of the previous experiments, in which changes in confidence must be inferred from aggregating data across trials. Moreover, I observed considerable variability in information seeking behaviour among participants. This variability restricted the usefulness of information seeking as a reliable proxy for confidence. Neural indicators of confidence are likely to be more stable across participants than information seeking indicators that consistently show wide individual differences that are largely unrelated to task performance (Desender et al., 2018; Pescetelli et al., 2021). Whereas some individuals consistently seek information on almost every trial and others rarely do so regardless of their confidence, such extremes are unlikely when it comes to neural markers. In this way, the experiment presented in this chapter extends the findings of the previous chapter, by now contrasting the effects of different social contexts on overtly reported confidence and continuous neural signatures of internal confidence.

The fundamental premise of this study hinges on the existence of a distinct neural correlate indicative of internal confidence. As outlined in the introductory chapter

of this thesis, neural substrates have been identified in which activity correlates with and lesions impair metacognitive judgements, such as confidence, but not object-level task performance. A particularly noteworthy study in relation to the experiment described in this chapter is an fMRI-based study conducted by Bang et al. (2020). This study employed an interactive task, requiring participants to adjust how they communicated their confidence about a perceptual decision based on the social context. The results revealed that whereas the dorsal anterior cingulate cortex (dACC) and perigenual anterior cingulate cortex (pgACC) tracked aspects of private confidence, the lateral frontal pole appeared to facilitate the context-dependent mapping of a private sense of confidence into a public report.

Notably, although Bang et al. (2020) employed fMRI and the ideas reviewed in the introductory chapter of this thesis are based on a multitude of neuroimaging techniques, evidence also supports the effective measurement of confidence using EEG. In fact, numerous EEG-based studies have discovered potential neural signatures linked with confidence (e.g., Balsdon et al., 2021; Desender, Murphy, et al., 2019; Gherman and Philiastides, 2018; Grogan et al., 2023; Herding et al., 2019; Kopčanová et al., 2023; Krumpe et al., 2019). For instance, a recent study by Dou et al. (2023) reported that a central-parietal EEG component encodes subjective confidence, independent of confounding factors such as reaction times, accuracy, and evidence quality. Additionally, a study by Sadras et al. (2022), which aimed to create a brain-computer interface (BCI) enhancing decision-making by supplying more information when necessary, employed multivariate decoding techniques and found that subjective confidence could be reliably decoded from single-trial, stimulus-locked, pre-response EEG data. This study holds particular relevance here, as I intend to utilise multivariate decoding of stimulus-locked pre-response EEG data to establish a marker of internal confidence, as will be described below. Furthermore, research by Desender, Murphy, et al. (2019) demonstrated that a classifier

trained to decode high versus low confidence from EEG data could predict information seeking choices, providing additional support to the proposition that an EEG marker indicative of internal confidence could align well with findings from the experiments reported in the previous chapter that employed information seeking to gauge internal confidence.

For the experiment described in this chapter, my initial approach to establishing an EEG marker of internal confidence was through multivariate decoding. This machine-learning method, increasingly acknowledged as a robust and efficacious tool for complex neural data analysis (Carlson et al., 2019), predicts or classifies outcomes based on patterns across multiple variables, rather than treating each variable in isolation. When applied to EEG data, multivariate decoding discerns patterns across multiple electrode channels and timepoints, decoding information manifested in distributed patterns of neural activity. The model is trained on a subset of the data, where it learns to associate neural activity patterns with varying states or conditions — in our case, levels of confidence. Upon training, the model is then applied to new, unseen data, making predictions based on its learned associations. In this study, the decoder's predictions will be used as a measure of internal confidence.

However, as will be detailed below, my efforts to decode confidence in this way met with limited success. This compelled a shift in strategy towards a more theory-guided approach, using the P3 event-related potential (ERP) as a marker of internal confidence. The choice of the P3 ERP as a potential marker was based on its established association with evidence accumulation and decision confidence, as corroborated by numerous studies (e.g., Fitzgerald et al., 2022; Rausch et al., 2020; Selimbeyoglu et al., 2012; Squires et al., 1975a, 1975b; Sutton et al., 1965; Sutton et al., 1982; Wilkinson and Seales, 1978; Zakrzewski et al., 2019). The P3 ERP is char-

acterised by a positive deflection in the EEG signal following stimulus presentation. This deflection typically manifests between 300 to 600 ms after the presentation of a stimulus, and is most pronounced at parietal and central electrode locations, with the highest amplitudes usually detected at the Pz electrode.

In a final, exploratory analysis, I took a more indirect analytic approach that focused on the Feedback-Related Negativity (FRN), an ERP associated with reward expectation and expectancy violation (Becker et al., 2014; Gehring and Willoughby, 2002; Holroyd and Coles, 2002; Holroyd et al., 2008). The FRN is marked by a relative negative shift for negative feedback compared to positive feedback at frontal-central recording sites. The maximum FRNs are typically detected at electrode FCz, peaking between 250 and 350 ms following feedback presentation. Crucially, previous research suggests that social misalignment, counter advice, or norm deviations can also trigger FRN responses in a way that is comparable to the way objective feedback does (Chen et al., 2012; Du et al., 2019; Kim et al., 2012; Lin et al., 2021). For example, in a line judgement task, Chen et al. (2012) observed that incongruent group choices triggered larger FRNs, with this response amplifying as the level of incongruity with group opinions increased. Similarly, Du et al. (2019), through an advice-taking study, found an increase in FRN amplitude corresponding to the distance of the advice given. Building on these findings, I explored whether the FRN would be more pronounced in disagreements involving high confidence partners compared to those with low confidence partners. This line of inquiry aligns with the identified correlation between social conflict magnitude and FRN amplitude, with high confidence partner disagreements potentially representing a larger social conflict. Furthermore, I probed whether this effect would be especially prominent in situations lacking objective feedback (the experimental context inducing informational conformity) as in these circumstances, the partner's confidence level could serve as a crucial performance evaluation indicator for the participant.

Using the same experimental setup as in the previous chapter, participants performed perceptual decision tasks under two contrasting conditions. The first condition presented a situation without feedback, but with exposure to a partner's choices and confidence reports (informational task condition). The second condition featured the presence of trial-wise feedback and a joint decision rule, wherein the decision reported with higher confidence dictated the pair's joint decision (normative task condition). EEG was recorded while participants completed the task, allowing me to examine how the relationship between EEG activity and participants' overt confidence reports varied under these different conditions and depending on the confidence bias of the partner they interacted with. In adapting the experiment to test internal confidence via EEG activity, I predicted that both conditions would still induce a matching of overt confidence reports, but with distinctive effects on the neural marker of confidence. This generates two primary predictions: (1) In the informational task condition (but not the normative one), the process of matching confidence with the partners' confidence reports should alter the neural markers of confidence, as measured via EEG; (2) In the normative task condition (but not the informational one), matching confidence with the partners' confidence reports should change the relationship between the EEG-based neural markers of confidence and the overt confidence reports. These predictions are illustrated in Figure 3.1, which shows the effects of the partners' influence on overtly reported confidence versus internal confidence. In this context, the '*internal confidence*' variable is no longer proxied by information seeking behaviour, but instead measured via EEG-activity related to confidence.

4.2 Methods

Participants. Participants were recruited via the Oxford Experimental Psychology Department's online participant recruitment system (SONA). Twenty-one partici-

pants completed the study. One participant was excluded prior to data analysis due to technical issues during EEG recording. This resulted in a final sample of 20 participants (10 female, mean age ± SD: 28.5 ± 7.1). Participants were compensated with a base amount of £10/hour plus an additional reward depending on performance in the tasks. The experiment was reviewed and approved by the University of Oxford's Medical Sciences Interdivisional Research Ethics Committee.

Procedure. The paradigm used for this experiment was similar to the one described in the previous chapter in which participants performed perceptual decision-making tasks in an informational and a normative setting (see Figure 3.2). As in Experiment 2 of the previous chapter, participants performed the dot discrimination task in one session and the colour discrimination task in the other session, with the order of the tasks counterbalanced across participants and task conditions (informational/normative). Critically, in contrast to the previous chapter, there were no information seeking trials because this study probes hypothesised neural rather than behavioural markers of internal confidence. Participants completed an initial 100 staircasing trials on their own (see previous chapter for details on the staircasing procedure), and then five blocks of 30 trials in one partner condition (overconfident/underconfident partner) followed by five blocks of 30 trials in the other partner condition (underconfident/overconfident partner).

EEG Acquisition. The experiment was conducted in an electrically shielded room. EEG data was recorded from 64 channels using a BioSemi ActiveTwo system (BioSemi, 2011). Electrode positions were Fp1, AF7, AF3, F1, F3, F5, F7, FT7, FC5, FC3, FC1, C1, C3, C5, TP7, CP5, CP3, CP1, P1, P3, P5, P7, P9, PO7, PO3, O1, Iz, I1, I2, Oz, POz, Pz, CPz, Fpz, Fp2, AF8, AF4, AFz, Fz, F2, F4, F6, F8, FT8, FC6, FC4, FC2, FCz, Cz, C2, C4, C6, TP8, CP6, CP4, CP2, P2, P4, P6, P8, P10, PO8, PO4, and O2. Horizontal and vertical electro-oculograms (EOG) were

recorded with electrodes placed above and below the left eye and at the outer canthi of both eyes to control for eye movement artefacts. Two additional electrodes were positioned on the mastoid bones behind the right and left ears. The Common Mode Sense (CMS) and Driven Right Leg (DRL) electrodes functioned as reference and ground electrodes. Data was continuously recorded at a sampling rate of 512 Hz.

EEG Preprocessing. EEG data preprocessing was performed with Python using the MNE software package (Gramfort et al., 2013) and custom code. Data was band-pass filtered between 0.5 and 30 Hz, and downsampled to 250 Hz for analysis. Data with large motor artefacts was removed by visual inspection. All electrodes were re-referenced to the mastoid electrodes. Independent component analysis (ICA) was then used to identify eye movement and blink artefacts from the EEG data. Eye-movement related ICA components were identified by visual inspection and by correlating their activity with that of the EOG channels, and removed from the EEG data. Stimulus-locked epochs were extracted beginning 300 ms prior to stimulus onset until 1000 ms post stimulus onset. Partner-response-locked epochs were extracted beginning 300 ms prior to the displaying of the partner's response until 1000 ms after. Epoch recordings were baseline-corrected to the average of the window -100 to 0 ms pre-stimulus and pre-partner-response, respectively. Epochs exceeding a peak-to-peak amplitude range of 200 µV were excluded.

Multivariate EEG Decoding. The aim of this study was to examine the relationship between internal and reported confidence in different social contexts. To this end, I first sought to identify a robust single-trial EEG signature of internal confidence via multivariate decoding. In common with recent EEG decoding studies of confidence (Fernandez-Vargas et al., 2020; Sadras et al., 2022; Valeriani et al., 2022), this analysis focused on the period between stimulus onset and the participant's response. Although other studies have also successfully decoded confidence

in a post-response period (as opposed to post-stimulus; e.g. Boldt and Yeung, 2015; Desender, Murphy, et al., 2019; Steinhauser and Yeung, 2010), the evidence suggests that activity in the post-response window scales mostly with error-detection in speeded tasks, whereas post-stimulus activity has been shown to scale with confidence in having made a correct decision (Boldt and Yeung, 2015). This is further substantiated by the fact that the P3 ERP, which has been associated with decision confidence in numerous studies (e.g. Fitzgerald et al., 2022; Rausch et al., 2020; Selimbeyoglu et al., 2012; Squires et al., 1975a, 1975b; Sutton et al., 1965; Sutton et al., 1982; Wilkinson and Seales, 1978; Zakrzewski et al., 2019), is a stimulus-locked signal. Notably, Sadras et al. (2022) exploring neural correlates of confidence using a task design that incorporated a post-stimulus gap to prevent the overlapping of stimulus-locked activity with response-locked activity, proposed that the neural correlates of confidence are stimulus-locked, and without a post-stimulus gap, these correlates could potentially be misinterpreted as response-locked. By controlling for response-related activity, it was shown that confidence could be accurately decoded from single-trial stimulus-locked pre-response EEG signals alone.

I tested whether single-trial confidence reports could be decoded from post-stimulus EEG activity, separately for each participant, across task conditions (i.e., across sessions completed on separate days) and across partner conditions. Multivariate decoding was performed using Ridge Regression with cross-validation, implemented through the Scikit-learn Machine Learning software package (Pedregosa et al., 2011). Ridge Regression is a linear regression method that incorporates a regularisation term to prevent overfitting and enhance generalisation to new data. For cross-validation, the data was divided into training and testing sets using 5-fold cross-validation, and the EEG feature data was normalised to have 0 mean and unit variance. In each cross-validation fold, a Ridge Regression model was trained trying each of 10 candidate regularisation strength parameter (α) values, which were evenly

spaced on a log scale between 10^{-5} and 10^{5}. A higher α value results in stronger regularisation, constraining the model by shrinking its coefficients, leading to a less complex model with lower variance but potentially higher bias. The optimal α value for each fold was selected based on which provided the best model performance. Models were trained to decode continuous confidence reports from EEG activity occurring between 200 ms and 800 ms post-stimulus.

The overall model performance for each participant was assessed by averaging the coefficients of determination, commonly known as the R-squared value, across the 5 folds. R-squared values typically fall between 0 and 1. These values represent the percentage of the dependent variable's (in this case, confidence reports) variance that can be predicted from the independent variables (in this case, the EEG activity). However, because here we are conducting out-of-sample prediction, when the model fit is exceptionally poor, the adjusted R-squared value can be less than 0. This happens when a predictive model's performance is so poor that it is worse than a baseline model that simply uses the mean of the observed outputs as the predicted output for all instances.

As will be detailed in the results section, various adaptations of this methodology were employed in pursuit of a reliable EEG correlate of confidence. Each unsuccessful attempt prompted subsequent alterations in the approach. The results section will detail these steps, illustrating how each failure instigated the next strategy. Ultimately, after exhaustive exploration, the decision was made to conclude that a reliable correlate of confidence was unattainable using the EEG decoding approach employed in this study.

Single-Trial ERP Analyses. Given that the decoding attempts yielded limited success in finding a robust single-trial measure of internal confidence, a more theory-driven ERP approach was adopted. I first investigated single-trial relation-

ships between confidence ratings and the stimulus-locked P3 ERP component. This approach focused on voltage at electrode Pz in the 1000 ms post stimulus onset, where the P3 component is typically observed. Then, I explored the relationship between partner agreement and the FRN ERP component, focusing on voltage at electrode FCz in the 1000 ms post partner response. Whereas the multivariate decoding approach tries to predict participants' confidence based on brain activity across EEG channels, the ERP analyses turn this relationship around, trying to predict brain activity at a specific EEG channel based on participants' confidence.

The ERP components were analysed using generalised linear models (GLMs), fitting an ordinary least squares regression to model the data. This step was implemented through the glmtools Python package (https://pypi.org/project/glmtools/). For each participant, separate GLMs with regressors of interest (in this case condition, partner, and reported confidence for the P3 analyses, and partner and agreement in each condition for the FRN analyses) were fitted for each timepoint at each channel, resulting in beta weights for every timepoint at every channel. To isolate the effects of each regressor and test for significance at the group-level, each regressor's beta coefficients were combined across participants, and then subjected to cluster-based permutation testing (5,000 permutations, $\alpha = .05$). This method circumvents the issue of multiple comparisons. During this procedure, a null distribution is generated for the channel of interest (in this case channel Pz for the P3 and channel FCz for the FRN) by randomly permuting the trial-average data at the participant level across the time dimension. Consequently, this allows for the modeling and testing of the observed group-level ERP against a 0 baseline.

4.3 Results

Behavioural Data. Participants' accuracy ranged from 53% to 78% with an average accuracy of 61%. This is below the targeted 72% accuracy, which was caused by a systematic technical error in the experiment code. Instead of setting the dot difference d for the task trials as the average of the smallest and the final d values achieved during the staircasing trials, I mistakenly set it as the average between those values on the logarithmic scale used for staircasing before translating the value on the logarithmic scale back into actual dot difference. This caused the dot difference to be skewed towards the smaller value (each participant's smallest d value achieved during the staircasing trials), leading to lower accuracy rates.

There was no consistent difference in participants' performance between the normative and informational task sessions or the underconfident and overconfident partner ($F_{task}(1, 19) = 0.65, p = .431$, $F_{partner}(1, 19) = 0.02, p = .896$, $F_{task \times partner}(1, 19) = 1.22, p = .282$). Participants' confidence reports were well-aligned with their objective accuracy, with higher confidence levels reported for correct answers compared to incorrect ones ($t(19) = 5.0, p < .001$ for the informational task and $t(19) = 3.9, p < .01$ for the normative task). Participants' metacognitive sensitivity was found to be above chance, as indicated by AUROC values significantly greater than 0.5 ($M = 0.60$, $SE = 0.05$, $t(19) = 9.7, p < .001$, when paired with the overconfident partner and $M = 0.61$, $SE = 0.06$, $t(19) = 7.8, p < .001$, when paired with the underconfident partner; see Figure 4.1). These values, though above chance, were generally lower than the AUROC values observed in the previous chapter's experiments ($M = 0.66$), which can be attributed to the higher task difficulty here.

Although there was no difference in participants' accuracy between partner condi-

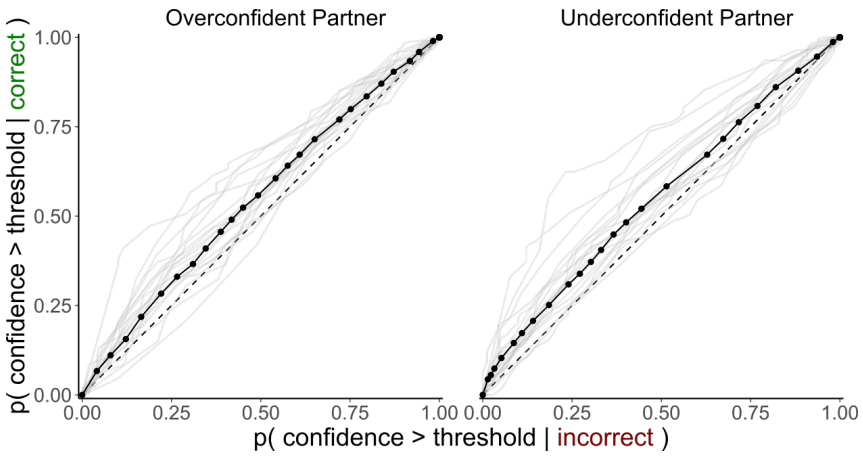


Figure 4.1: Metacognitive performance with each partner. Faint lines show individual participants' Receiver Operator Characteristic (ROC) curves. Solid lines show mean data across participants. ROC curves plot for each confidence value the probability of a confidence report being above that value for correct responses against the probability of a confidence report being above that value for incorrect responses. The area under the ROC curve (AUROC) is a measure of how well confidence ratings distinguish between correct and incorrect responses, whereby higher AUROC means higher metacognitive sensitivity with a value of 0.5 indicating performance at chance level (dotted line) and an AUROC of 1 indicating perfect discrimination.

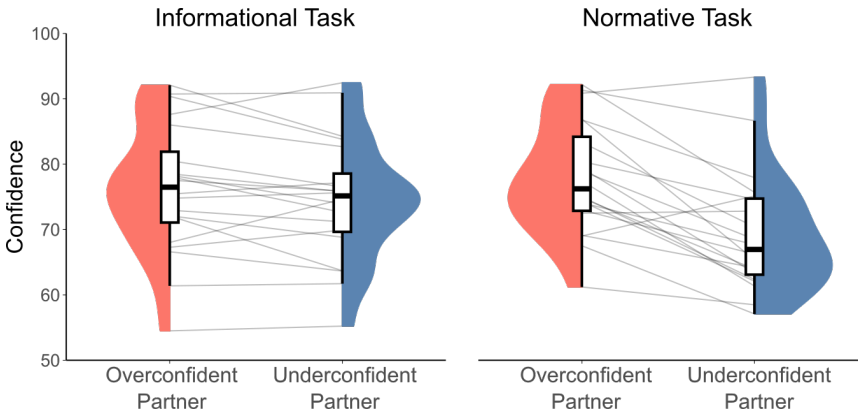


Figure 4.2: Confidence in the informational and normative tasks when paired with the underconfident and overconfident partner. Faint lines show individual participant means. Violin and box plots show the distributions of participant means.

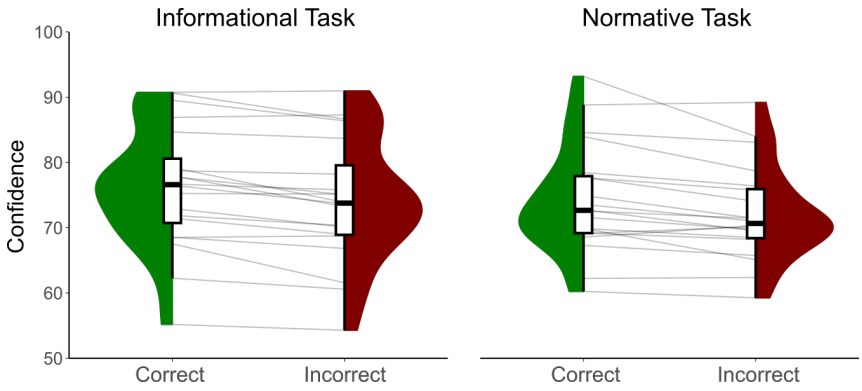


Figure 4.3: Confidence in the informational and normative tasks for correct and incorrect responses. Faint lines show individual participant means. Violin and box plots show the distributions of participant means.

tions ($t(19) = 0.1, p = .90$), their confidence levels significantly differed (see Figure 4.2): As predicted, participants reported higher confidence when paired with the overconfident partner, and lower confidence when paired with the underconfident partner. An ANOVA examining the effects of task condition (normative versus informational), partner condition (overconfident versus underconfident), and their interaction revealed a significant effect of partner condition on participants' reported confidence and a significant interaction between task and partner ($F_{task}(1, 19) = 0.83, p = .37$, $F_{partner}(1, 19) = 24.5, p < .001$, $F_{task \times partner}(1, 19) = 22.9, p < .001$). Follow-up t-tests indicated a significant difference in confidence between partners for the normative task ($t(19) = 5.6, p < .01$), but the difference in confidence for the informational task was only significant in a one-side test, which, however, is warranted given my a priori hypothesis about the direction of the effect ($t(19) = 1.9, p = .07$ two-sided, $p < .05$ one-sided). Compared to the experiments in the previous chapter, the observed confidence shifts in the current experiment were similar for the normative task condition ($M_{shift} = 8.5$, $SE = 6.8$ compared to 7.5 in Experiment 1 and 7.6 in Experiment 2) but smaller in the informational task condition ($M_{shift} = 1.7$, $SE = 3.9$ compared to $M = 4.6$ in Experiment 1 and $M = 5.7$ in Experiment 2). The shift in confidence between correct and incorrect responses ($M_{shift} = 2.5$,

$SE = 2.2$) was also smaller than that in the previous experiments ($M = 4.7$ confidence points in Experiment 1 and $M = 5.2$ confidence points in Experiment 2; see Figure 4.3). Of interest was how these shifts in reported confidence between partner conditions would relate to EEG correlates of confidence, and whether this relationship differed across task conditions (informational versus normative).

Multivariate EEG Decoding.

Spatiotemporal EEG decoding: timepoints × channels. To establish an EEG correlate of confidence, I employed multivariate decoding. The initial analysis incorporated both EEG channel and timepoint dimensions as features in the Ridge Regression models, intending to exploit all accessible spatial and temporal information to predict confidence reports. This approach generated positive R-squared values for 14 out of 20 participants (see Figure 4.4).

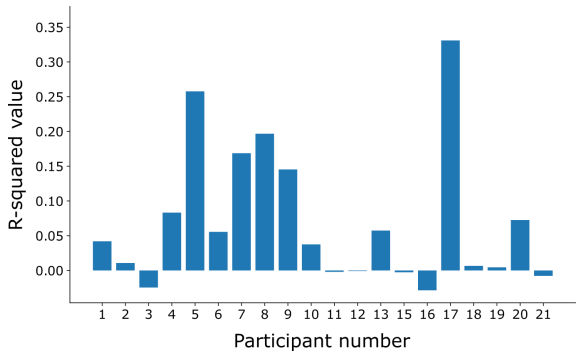


Figure 4.4: Performance of the Ridge Regression models that employ channels × timepoints as features to predict participants' trial-wise confidence. Each bar represents a participant. The bar plot shows the R-squared values averaged across the 5 cross-validation folds, indicating the proportion of variance in each participant's confidence reports that the models can account for. Higher R-squared values indicate better predictive performance.

For these 14 participants, the average R-squared value was 0.105. Figure 4.5 demonstrates, for one of the 5 folds, the model performance in the training data (first column) and the test data (second column) by plotting actual confidence reports on the

x-axis against decoder predicted confidence reports on the y-axis for each participant. Figure 4.5 also shows how model performance varies between the 5 folds (third column), plotting the Mean Squared Error for training data (blue) and test data (orange). The plot suggests that the predicted confidence is heavily biased towards each participant's mean confidence and shows little variation, possibly explaining the relatively poor model performance as indicated by the low (for 6 participants even < 0) R-squared values. This might be caused by over-strong regularisation. However, given that the α value was chosen in a range of 10^{-5} to 10^{5}, a smaller α value would have been selected, had it performed better.

Moreover, in a preceding effort with less regularisation, significant overfitting issues were encountered. Overfitting is a challenge inherently posed by high-dimensional problems, a category into which my Ridge Regression problem falls due to a higher number of features (150 timepoints $\times$ 64 channels) compared to observations (600 trials). In such high-dimensional problems, the risk of overfitting amplifies, since the model may perfectly adapt to data noise rather than capturing the underlying trend. In this preceding attempt with less regularisation, evident from Figure 4.6, the model showed a high predictive performance in training data, as indicated by the close alignment between actual and predicted confidence in the first column's scatter-plots, but failed to generalise this performance onto test data, as evident from the lack of alignment between actual and predicted confidence in the second column' scatter-plots, indicating overfitting. This mismatch is further confirmed by the significantly larger Mean Squared Error for the test data (orange) compared to the train data (blue; third column), suggesting that the model was overly tuned to the specifics of the training data, including its noise and outliers, leading to poor performance when presented with new, unseen data.

Dimension reduction by averaging across timepoints. To mitigate these overfitting

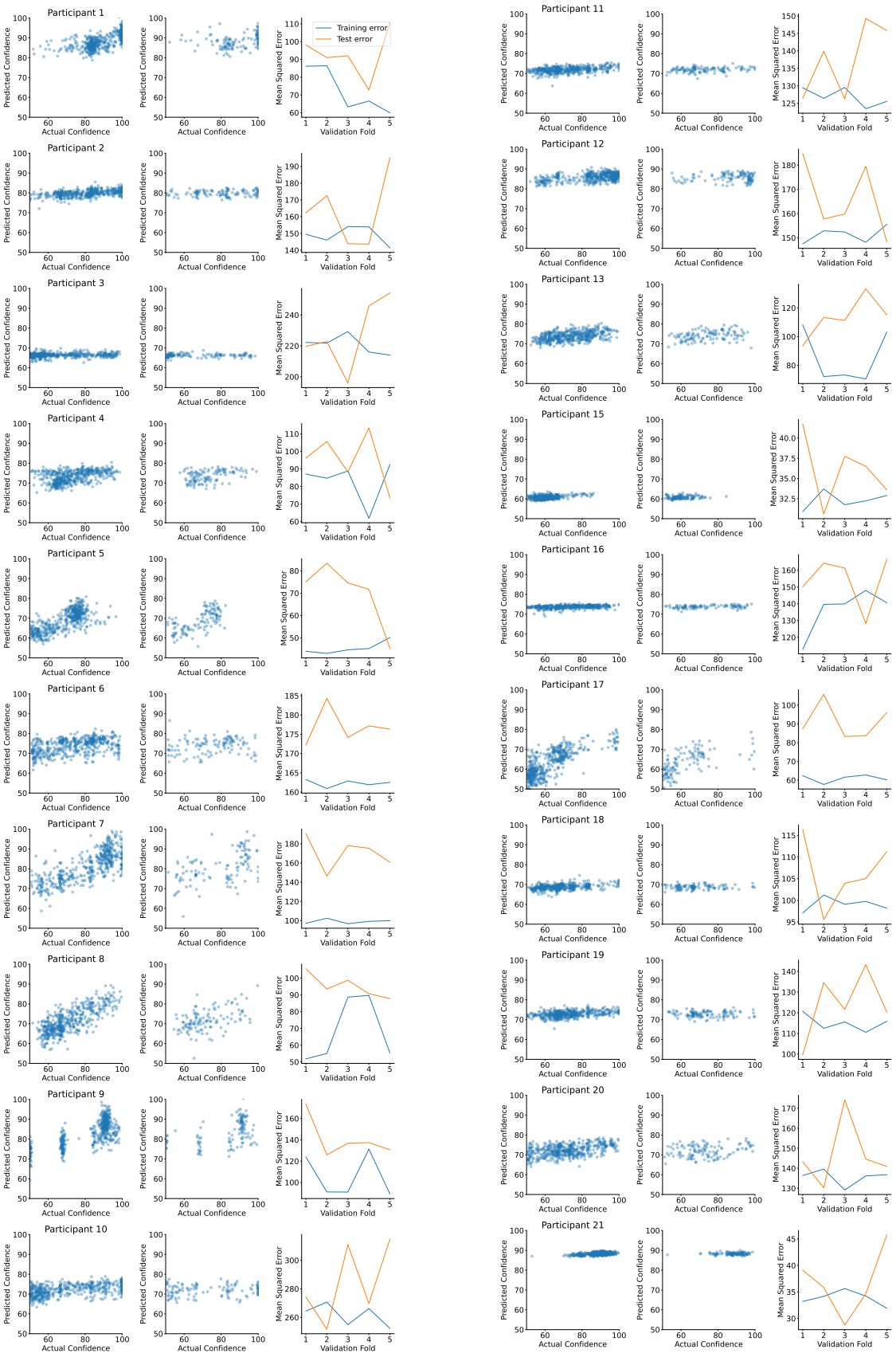


Figure 4.5: Performance of the Ridge Regression models that employ channels × timepoints as features to predict participants' trial-wise confidence. Individual subplots per participant show: (1) actual versus predicted confidence in training data for the 5th fold of cross-validation, (2) actual versus predicted confidence in test data for the 5th fold of cross-validation, and (3) Mean Squared Error for training and test data over the 5 cross-validation folds.

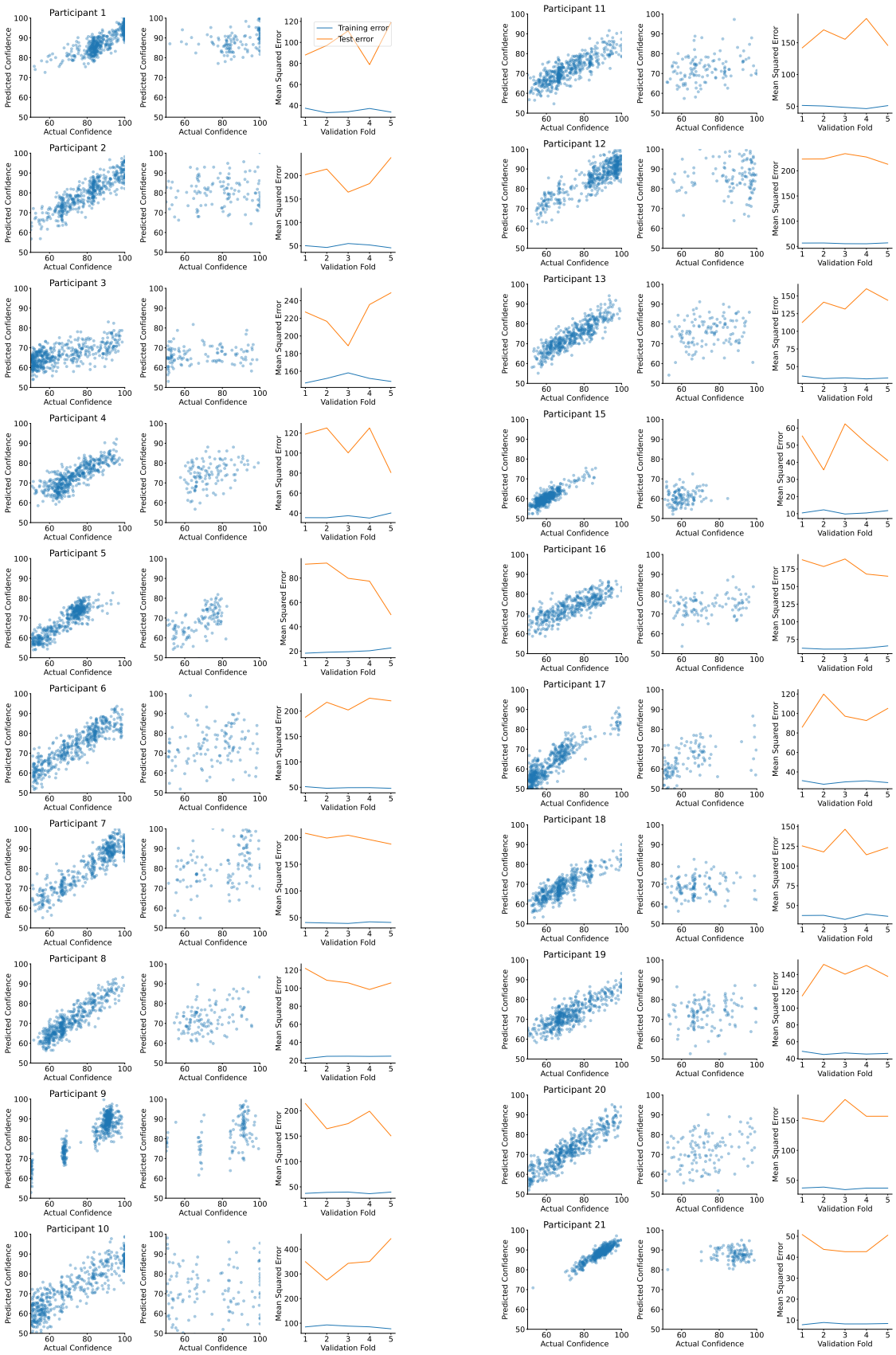


Figure 4.6: Performance of the Ridge Regression models that employ channels × timepoints as features to predict participants' trial-wise confidence with less regularisation. Individual subplots per participant show: (1) actual versus predicted confidence in training data for the 5th fold of cross-validation, (2) actual versus predicted confidence in test data for the 5th fold of cross-validation, and (3) Mean Squared Error for training and test data over the 5 cross-validation folds.

issues associated with high dimensionality when utilising channel × timepoints as model features, a revised strategy was employed to predict participants' confidence from their EEG data. This updated approach sought to reduce the dimensionality in the EEG data by averaging across timepoints, thereby reducing the features to the number of channels. All other modeling steps remained unchanged. However, this approach yielded R-squared values greater than 0 for only 9 participants, with an average R-squared value of 0.075 for these participants (see Figure 4.7).

Dimension reduction via Principal Component Analysis. In another effort to decrease the dimensionality of the EEG data without sacrificing temporal information as in the previous approach, Principal Component Analysis (PCA) was utilised. PCA was applied to the normalised EEG data to reduce its dimensionality, preserving 95% of the variance in the data. On average, 125 components were extracted for each participant (see Figure 4.8A). The transformed data was then used to train and test the prediction model as before. This method yielded R-squared values greater than 0 for 14 participants, with an average R-squared value of 0.104 for these individuals (see Figure 4.8B). There was no clear association between the number of components extracted and the decoding performance.

Timepoint-wise decoding. In a last approach, Ridge Regression was applied to each timepoint separately to explore whether the model's predictive performance improved at specific timepoints. For every participant, a Ridge Regression was conducted utilising a 5-fold cross-validation at each timepoint. As in the previous approaches, this process included the selection of the regularisation (α) parameter that maximized performance, tested among 10 values on a logarithmic scale ranging from $10^{-}5$ to 10^{5}. The first column in Figure 4.9 represents the progression of R-squared values, averaged across the 5 folds, over time post-stimulus for each participant. I identified the timepoint with the maximum R-squared value for ev-

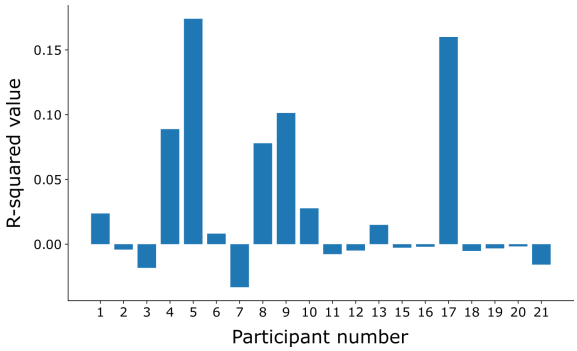


Figure 4.7: Performance of the Ridge Regression models that use the average voltage across timepoints at each channel to predict participants' trial-wise confidence. Each bar represents a participant. The bar plot shows the R-squared values averaged across the 5 cross-validation folds, indicating the proportion of variance in the participant's confidence reports that the model can account for. Higher R-squared values indicate better predictive performance.

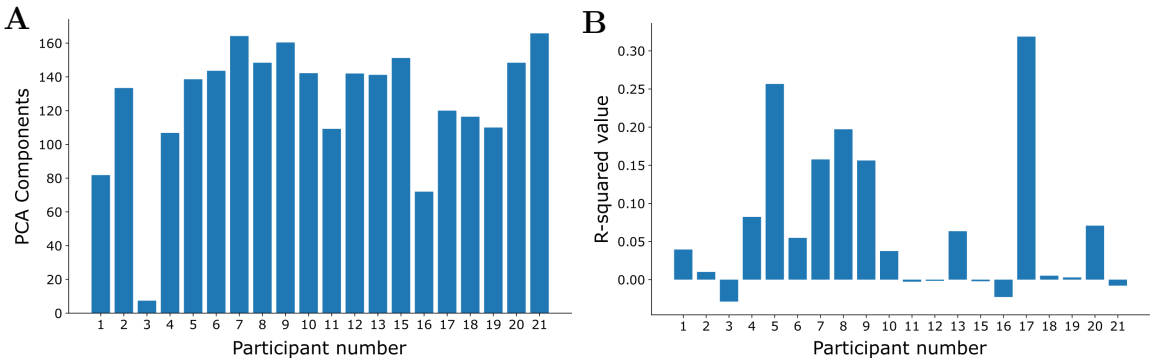


Figure 4.8: **(A)** Average number of PCA components retained across 5 cross-validation folds for each participant. Components were selected to retain 95% of the variance in the EEG data. **(B)** Performance of the Ridge Regression models that employ PCA components as features to predict each participant's confidence. Each bar represents a participant. The bar plot shows the R-squared values averaged across the 5 cross-validation folds, indicating the proportion of variance in the participant's confidence reports that the model can account for. Higher R-squared values indicate better predictive performance.

ery participant. This approach proved more successful compared to earlier methods, with the R-squared value at each participant's respective '*best*' timepoint registering positive values for 18 out of 20 participants. The average of these R-squared values for the 18 participants was 0.072, with this best decoding performance observed on average 522 $\pm$ 168 milliseconds following stimulus onset.

The second column of Figure 4.9 presents topoplots, which offer a visual representation of the beta weights, averaged over the 5 folds, for the channels at each participant's '*best*' timepoint. These topoplots appear patchy, reflecting the fact that activity at neighbouring channels is necessarily highly correlated such that the model is forced to attribute the variance in that activity between the channels in the regression model. In an effort to better comprehend the neuroanatomical basis of the predictive patterns, sensor projections were calculated and are illustrated in Figure 4.9's third column. These projections aim to illustrate the EEG channel activity correlating with the distinguishing component, as per Parra et al. (2002). However, upon comparing the sensor projection patterns across different participants, there is a striking lack of consistency. If the distinguishing component is indeed linked to a confidence computation process, one might assume a relatively uniform manifestation of this process across participants, given that it is considered to be a widely shared cognitive function, not unique to each individual. This observation might imply that the observed distinguishing component captures more than just the neural processes directly linked to confidence computation. It might also include noise or other neural activities that are unrelated to confidence but differ from individual to individual.The differences could stem from the high dimensionality and inherent complexity of the EEG data, which may result in models that are overly fitted to individual-specific noise rather than capturing the generalisable trend of confidence computation.

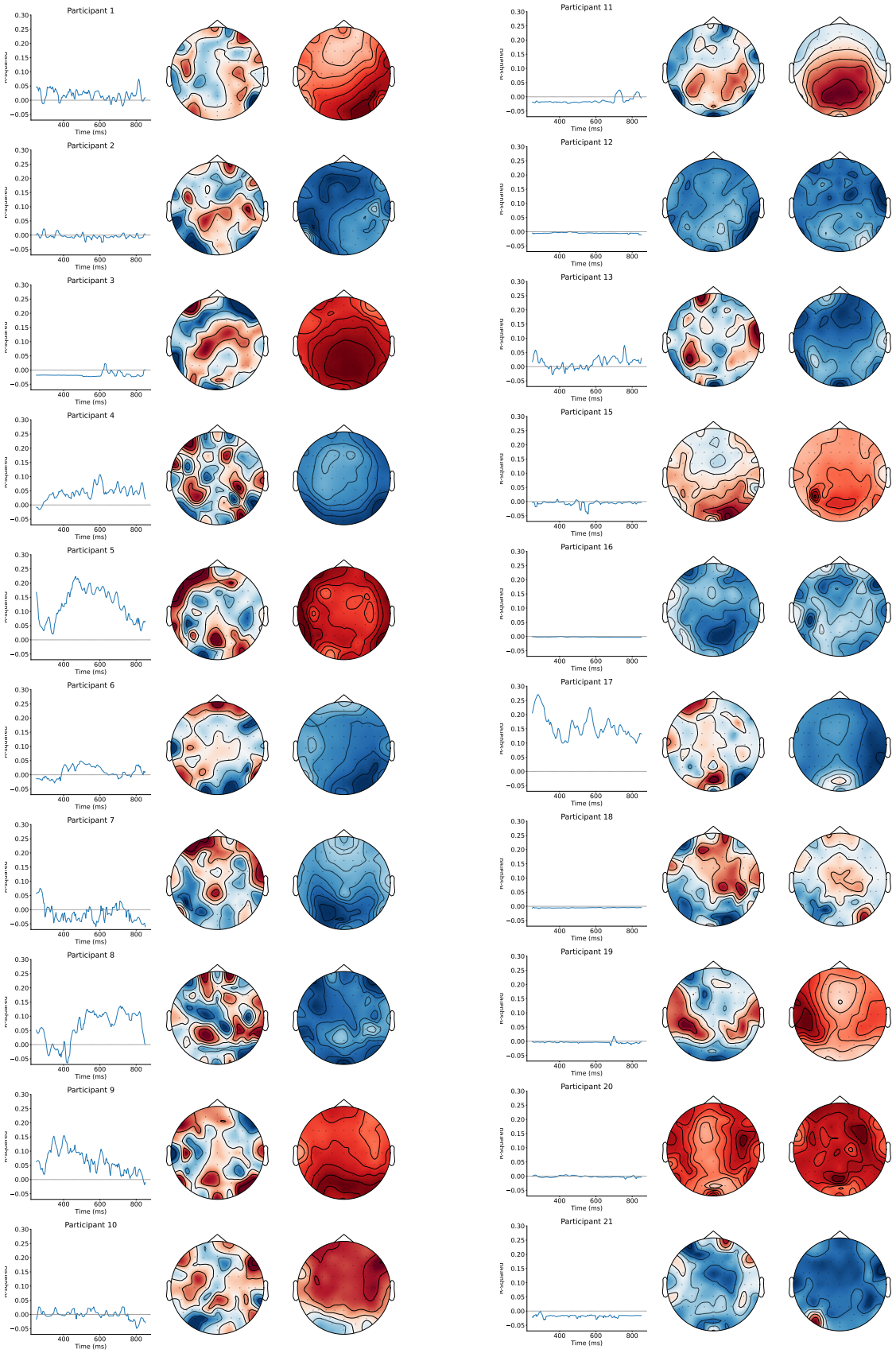


Figure 4.9: Time-point wise decoding of participant confidence from EEG data using Ridge regression. Individual subplots per participant show: (1) timecourse of R-squared values, indicating the accuracy of the model over time, (2) topoplot of channel beta weights at the point of maximum R-squared value, and (3) topoplot of sensor projections at the point of maximum R-squared value.

I derived decoder-based estimates of confidence for each participant at their identified '*best*' timepoint. In each fold, decoder-based estimates were obtained for the test data based on the training data in that fold. Despite these decoded values certainly not providing a perfect measure of confidence, given the relatively low predictive performance of the models, the cross-validation approach employed when generating these values ensures that they are not merely reflecting overfitting or idiosyncratic noise in the data, but rather capture some consistent, reproducible aspect of the neural data that is associated with confidence. Therfore, excluding the two participants with R-squared values below 0, these decoded confidence values were subsequently used as an '*internal confidence*' metric to test the main hypotheses.

Hypothesis testing. Given the mostly successful induction of confidence matching in both experimental contexts with (assumed) informational and normative influences, the critical analyses examined whether the observed effects of partner confidence on participants' overt confidence reports diverged from effects on their internal confidence as indexed by the decoder-based estimates of participants' confidence. Firstly, to test the hypothesis that informational and normative influences differently impact internal confidence, I tested the effects of overconfident and underconfident partners on participants' decoded confidence in both informational and normative task conditions. Figure 4.10 visually presents the results, which were analysed via an ANOVA to evaluate the effect of task condition, partner condition, and their interaction on decoded confidence. My primary prediction was an interaction between task and partner conditions. Specifically, if informational influence alters internal confidence, but normative influence does not, a shift in decoded confidence between partner conditions should solely appear in the informational task condition. However, the ANOVA results showed that there was no significant variation in decoded confidence based on either partner condition ($F(1, 17) = 0.2, p = .67$), task condition ($F(1, 17) = 1.3, p = .27$), or their interaction ($F(1, 17) = 1.4, p = .25$). Thus, the

results did not support my hypothesis; decoded confidence did not change depending on the partner's confidence bias in the informational task.

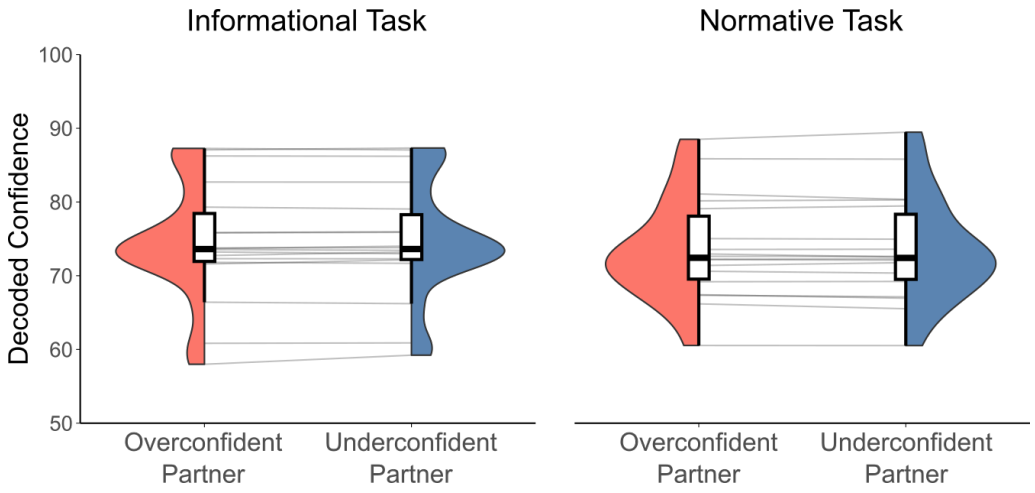


Figure 4.10: Decoded confidence in the informational and normative tasks when paired with the underconfident and overconfident partner. Faint lines show individual participant means. Violin and box plots show the distributions of participant means.

In the second key analysis, I assessed the correlation between reported confidence and decoded confidence on a trial-by-trial level. The objective was to probe the hypothesised relationship and dissociation between these measures across partner conditions, as depicted in Figure 3.1. Figure 4.11 presents the consolidated results, with the y-axis representing participants' decoded confidence, serving as a proxy for internal confidence, and the x-axis representing participants' overtly reported confidence. A linear mixed-effects model (LMM) was applied, predicting trial-wise decoded confidence by reported confidence, task condition, partner condition, and an interaction term between task and partner conditions, with random intercepts and slopes per participant. Reported confidence was significantly associated with decoded confidence ($\beta = 0.027, SE = 0.01, \chi^2 = 20.92, p < .001$), which was expected given that the decoded confidence values were derived from the reported confidence. However, none of the other predictors significantly influenced decoded confidence

(partner: $\beta = 0.130, SE = 0.09, \chi^2 = 3.15, p = .076$, task: $\beta = 0.937, SE = 0.81, \chi^2 = 1.55, p = .213$, interaction: $\beta = 0.014, SE = 0.12, \chi^2 = 0.01, p = .911$).

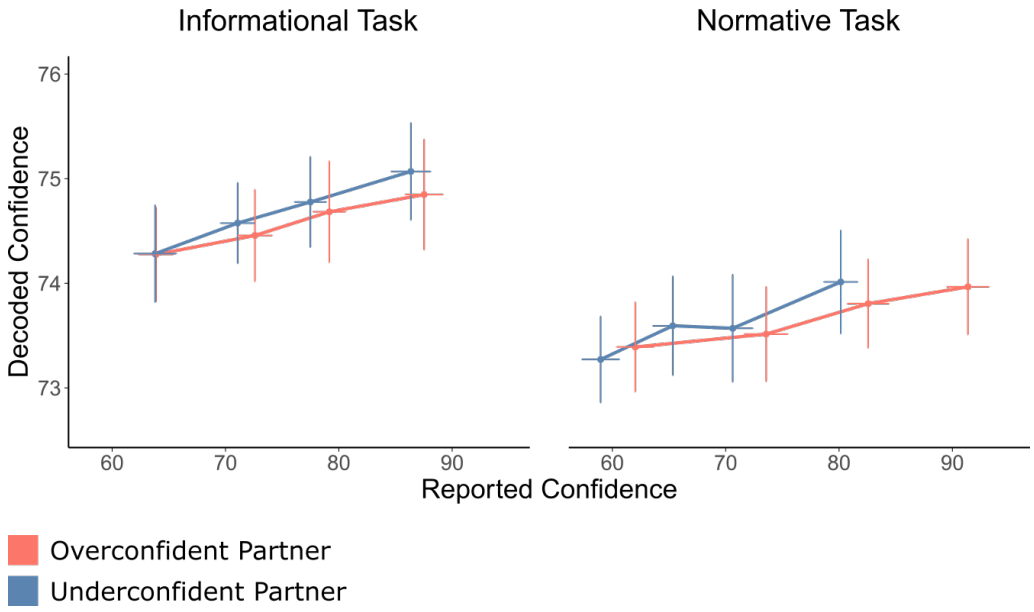


Figure 4.11: Decoded confidence as a function of reported confidence in the informational and normative tasks when paired with the underconfident and overconfident partner. For each participant, reported confidence was divided into quartiles and the mean decoded confidence within each reported confidence quartile was computed. The graph displays the grand mean across participants and standard error of the mean for both reported confidence and decoded confidence.

The null-results seem to reflect the fact that the variation in decoded confidence values across conditions was minimal, with the data primarily clustering around each participant's average confidence level. This is graphically evident in Figure 4.11, in which decoded confidence does scale with reported confidence but note that the y-axis range is very small. Similarly, this is evident from the small beta-value ($\beta = 0.027$) for the relationship between reported and decoded confidence in the LMM above. This lack of variance in decoded confidence values could be attributed to a bias introduced through strong regularisation applied when deriving the decoder-based confidence values. However, a decrease in the degree of regularisation raises the risk of overfitting to the training data, a potential pitfall that was evidenced

earlier. In light of this, the absence of an effect here likely reflects problems in identifying a reliable proxy for internal confidence in the first place, rather than an absence of the hypothesised effects in general.

Given these constraints and the inconclusive findings from the decoding analyses, I turned to a more theory-driven approach for further investigation, wherein I investigated ERPs, as described next.

Single-Trial P3 Data. Figure 4.12 shows the grand average voltage for each channel in the stimulus epoch. As can be seen in Figure 4.12, there is a positive deflection in the EEG signal at parietal and central electrode locations between 300 and 600 ms after stimulus presentation, looking like a P3 response. Figure 4.13 displays the grand-average ERP at electrode Pz, where the highest P3 amplitudes are usually detected. In the informational task, the P3 appears to scale with reported confidence, and does so regardless of the partner condition. This is evident from the fact that the solid lines, showing the ERP for trials with higher than median confidence, are generally higher than the dotted lines, showing the ERP for trials with lower than median confidence (see Figure 4.13 A). In contrast, in the normative task, the P3 only appears to scale with confidence when participants interacted with the underconfident partner (see Figure 4.13 B).

The P3 ERP component was analysed using GLMs with a design matrix containing a constant regressor to model the intercept, a parametric regressor for z-transformed confidence, and categorical regressors for both the task condition (informational versus normative) and the partner condition (underconfident versus overconfident) as well as their interaction. Figure 4.14 shows an exemplary design matrix for one participant. In line with the analyses conducted for the experiments in the previous chapter, I did not include an interaction term between confidence and task condition or partner condition. This is because the hypothesised effect is that there will be

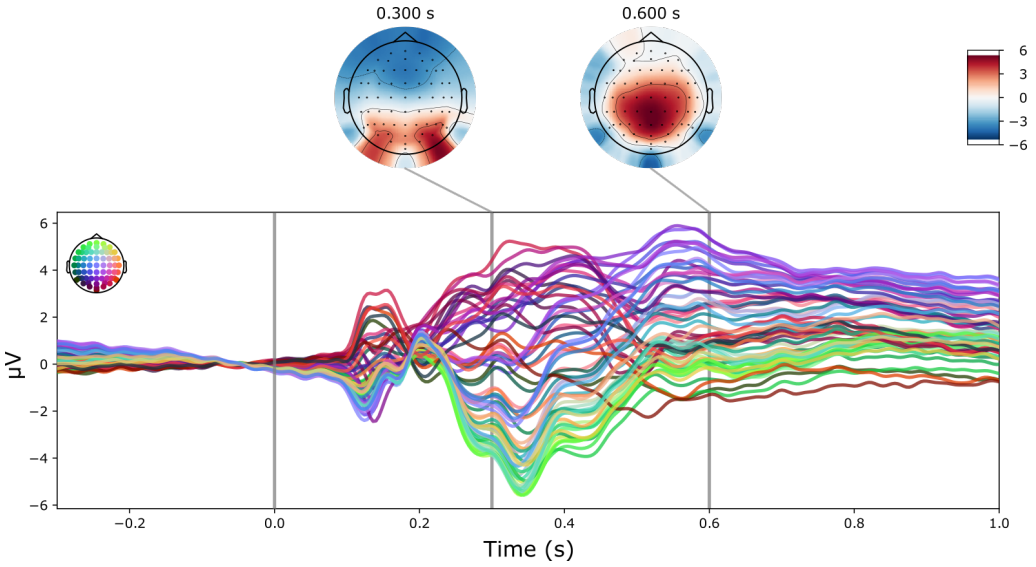


Figure 4.12: Grand average voltage for each channel in the stimulus epoch.

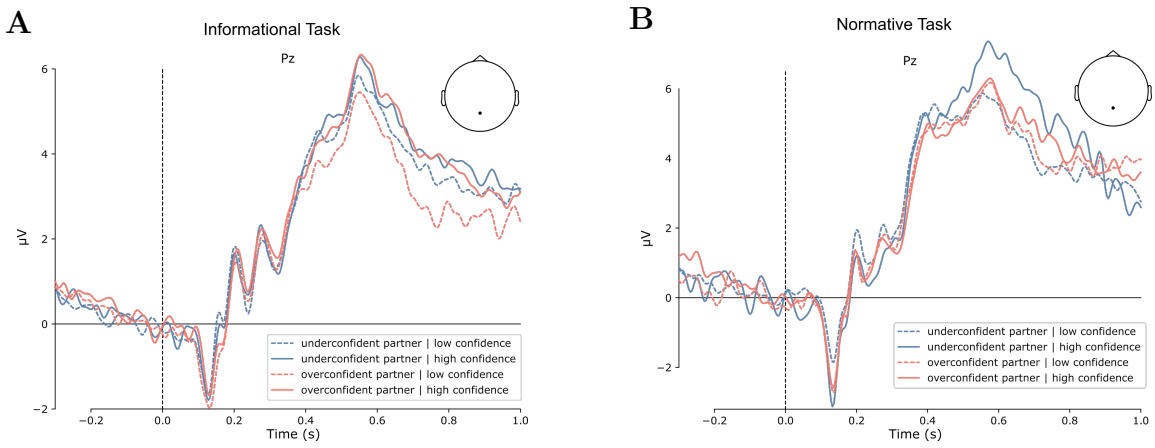
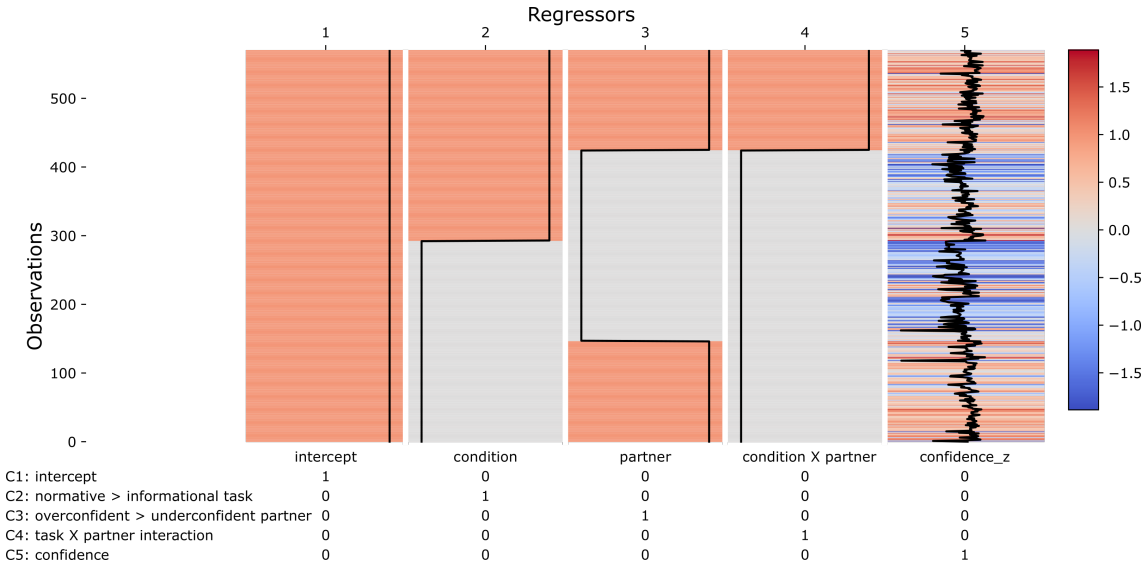


Figure 4.13: Grand-average ERP at electrode Pz in the **(A)** informational and **(B)** normative task conditions. Time is relative to stimulus onset. Confidence is median split across both partner conditions.

an offset rather than a change in slope between internal and reported confidence between partner conditions within the normative task condition.



C1: intercept	1	0	0	0	0
C2: normative > informational task	0	1	0	0	0
C3: overconfident > underconfident partner	0	0	1	0	0
C4: task X partner interaction	0	0	0	1	0
C5: confidence	0	0	0	0	1

Figure 4.14: Example design matrix for a single participant with individual regressors in columns. The table underneath shows the contrast matrix with corresponding weightings for each regressor. The intercept was defined as the informational task condition with the underconfident partner.

Figure 4.16 shows the model-derived ERPs. The intercept models the ERP for the informational task condition with the underconfident partner. Whilst controlling for task and partner condition, a significant relationship was observed between confidence reports and a late component of the evoked potential, with higher amplitudes of the ERP preceding higher confidence ratings (see Figure 4.16 C5). A significant cluster for the effect of confidence on ERP amplitude emerged from 532 to 840 ms post stimulus onset. However, no significant effects were found for task condition (C2), partner condition (C3), or their interaction (C4) on ERP amplitude. Although there were no significant effects of task and partner conditions, a separate GLM that did not control for these factors but used only confidence reports and experiment block count as regressors did not yield a significant effect of confidence on ERP amplitude. This set of findings suggests that task and partner conditions, although

not exerting a direct significant effect on ERP amplitude, may still play a role by explaining some meaningful variance in the relationship between overtly reported confidence and P3 amplitude — a potential index of internal confidence.

Single-Trial FRN Data. For the final exploratory analysis, my objective was to investigate whether disagreement versus agreement of a partner elicited differing brain responses based on whether the interaction was with a low or high confidence partner. In particular, I was interested in determining whether this potential variation itself differed between the informational and normative task conditions. The central hypothesis revolved around the notion that participants might interpret partners' confidence and agreement differently depending on the task condition (informational/normative), and that this might manifest in the neural responses invoked upon the unveiling of the partner's response. To address these questions, I focused on examining the FRN ERP at electrode FCz. As illustrated in Figure 4.15, the ERP following a partner's response varied depending on whether the partner's response agreed or disagreed with the participant's choice in both the informational and normative tasks. This is evident from the fact that the solid lines, showing the ERP for agreement trials, are generally higher than the dotted lines, showing the ERP for disagreement trials. However, this difference between agreement and disagreement does not resemble a traditional FRN, but rather illustrates a sustained difference in voltage. Only in the case of interacting with the overconfident partner in the normative task condition does the variation appear similar to an FRN. This can be observed from the dotted pink line representing the ERP for disagreement trials, which exhibits a negative deflection around 300 ms following the partner's response, compared to the solid pink line, representing the ERP for agreement trials.

To investigate statistical differences, separately for each task condition, I constructed GLMs with a design matrix containing a constant regressor to model the intercept,

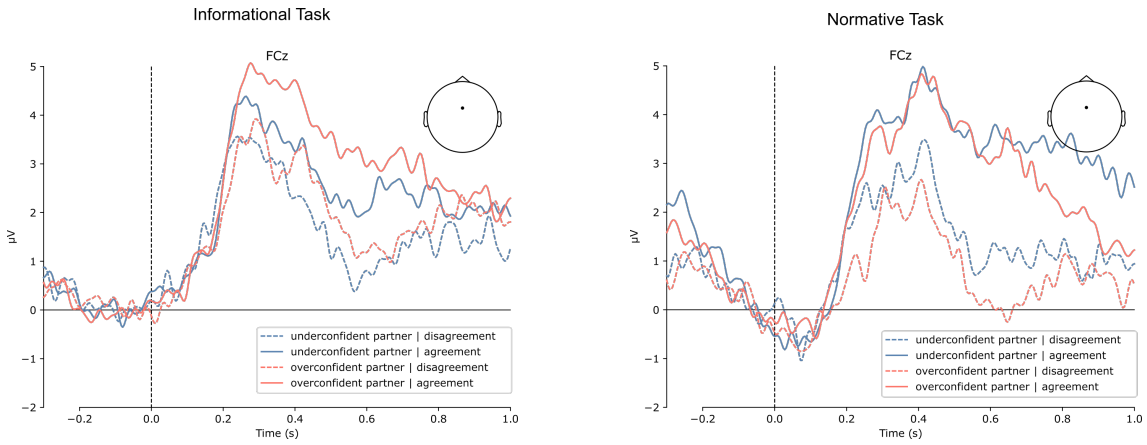


Figure 4.15: Grand-average ERP at electrode FCz in the informational and normative tasks. Time is relative to stimulus onset.

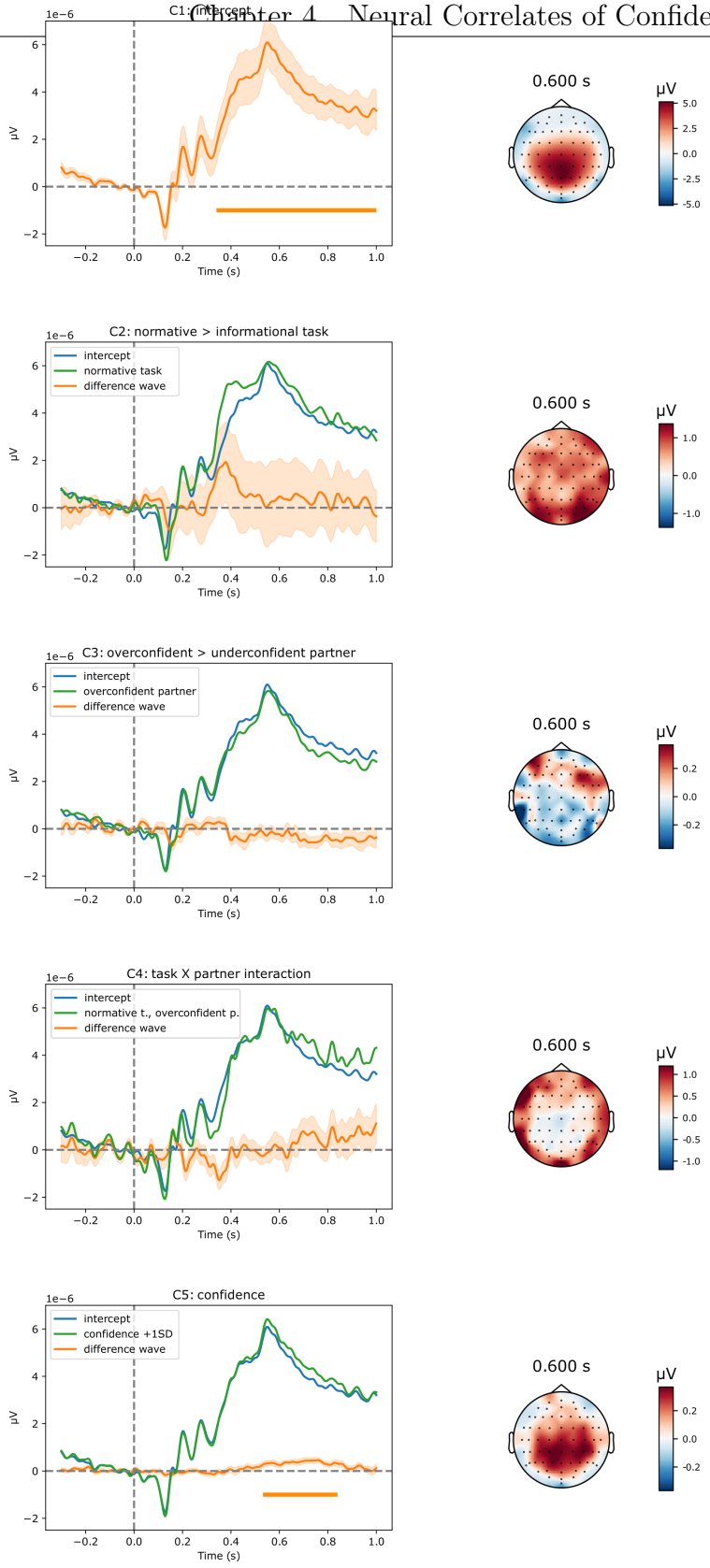


Figure 4.16: Modelled grand-average ERP at electrode Pz and visualisation of scalp topographies for the difference wave at 600 ms following stimulus presentation. The ERP traces depict the mean beta coefficient across participants for the relevant condition or contrast. The intercept models the ERP for the informational task condition with the underconfident partner. Shaded areas represent the standard error of the mean. Orange bars below the x-axis denote time periods of the difference contrasts that survived permutation cluster testing against 0. C stands for contrast.

and categorical regressors for the partner condition (underconfident versus overconfident) and the partner's agreement (agreement versus disagreement), as well as their interaction. Figure 4.17 and Figure 4.18 show the model-derived ERPs. The intercept models the ERP for disagreement trials with the underconfident partner. In the informational task, a significant cluster for the effect of agreement on ERP amplitude emerged from 540 to 700 ms post partner response (see Figure 4.17 C3). In the normative task, a significant cluster for the effect of agreement on ERP amplitude emerged from 436 to 920 ms post partner response (see Figure 4.18 C3). However, there were no main effects of partner condition as well as no interaction effects of agreement with partner condition in either of the task conditions. The absence of an interaction suggests that, whereas the disagreement of a partner affected the ERP in both informational and normative tasks, against predictions, this agreement effect was not modulated by the partner's confidence level in either task condition.

4.4 Discussion

The experiment reported in this chapter was designed as an extension of the investigations in the preceding chapter, aiming to further characterise the mechanisms underpinning confidence matching in social contexts. It focused on the roles of informational and normative conformity, specifically examining how the latter may lead to a discrepancy between overtly reported confidence and internal confidence. In comparison to the previous experiments, which gauged internal confidence through participants' information seeking choices, this study focused on neural correlates of confidence, leveraging EEG to record participants' brain activity during the performance of a perceptual decision-making task. In line with the previous chapter, the experimental design comprised two conditions: an informational task, in which participants observed a partner's decisions and confidence ratings without receiving

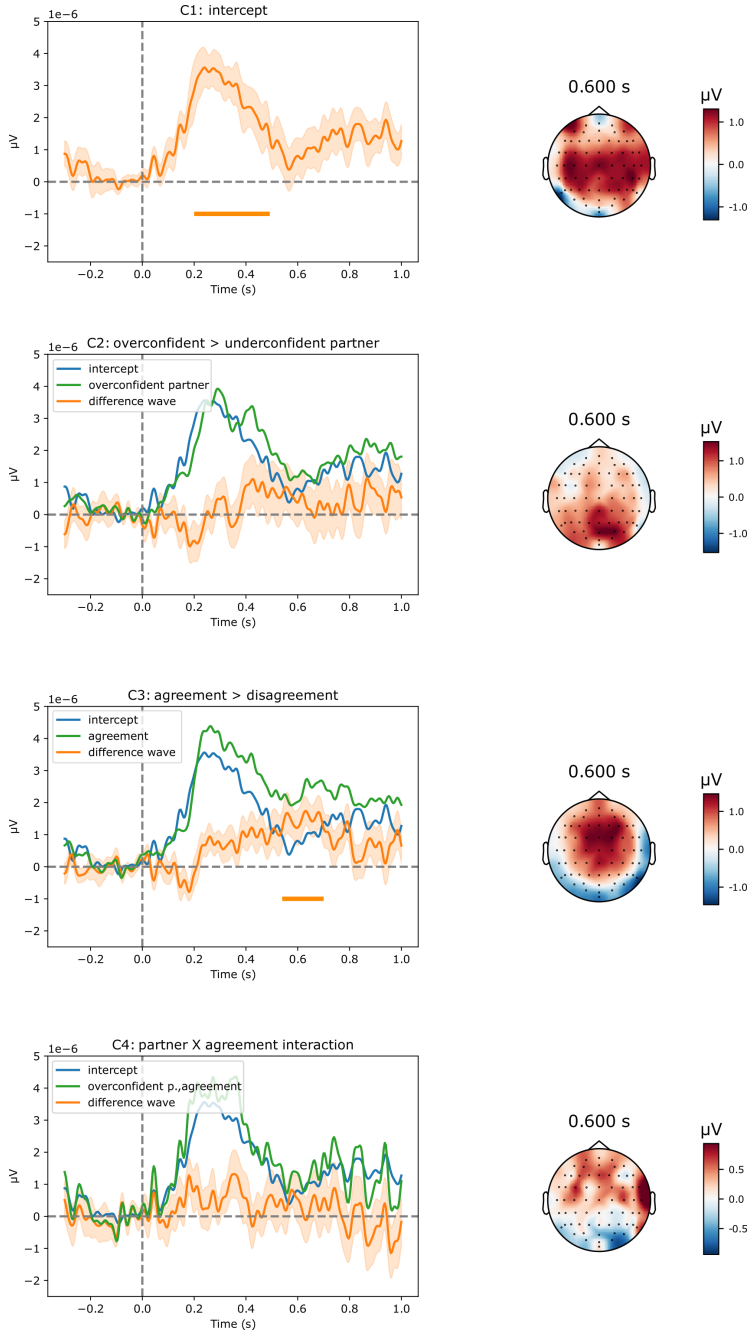


Figure 4.17: Modelled grand-average ERP at electrode FCz and visualisation of scalp topographies for the difference wave at 600 ms following stimulus presentation for the informational task condition. The ERP traces depict the mean beta coefficient across participants for the relevant condition or contrast. The intercept models the ERP for disagreement trials with the underconfident partner. Shaded areas represent the standard error of the mean. Orange bars below the x-axis denote time periods of the difference contrasts that survived permutation cluster testing against 0. C stands for contrast.

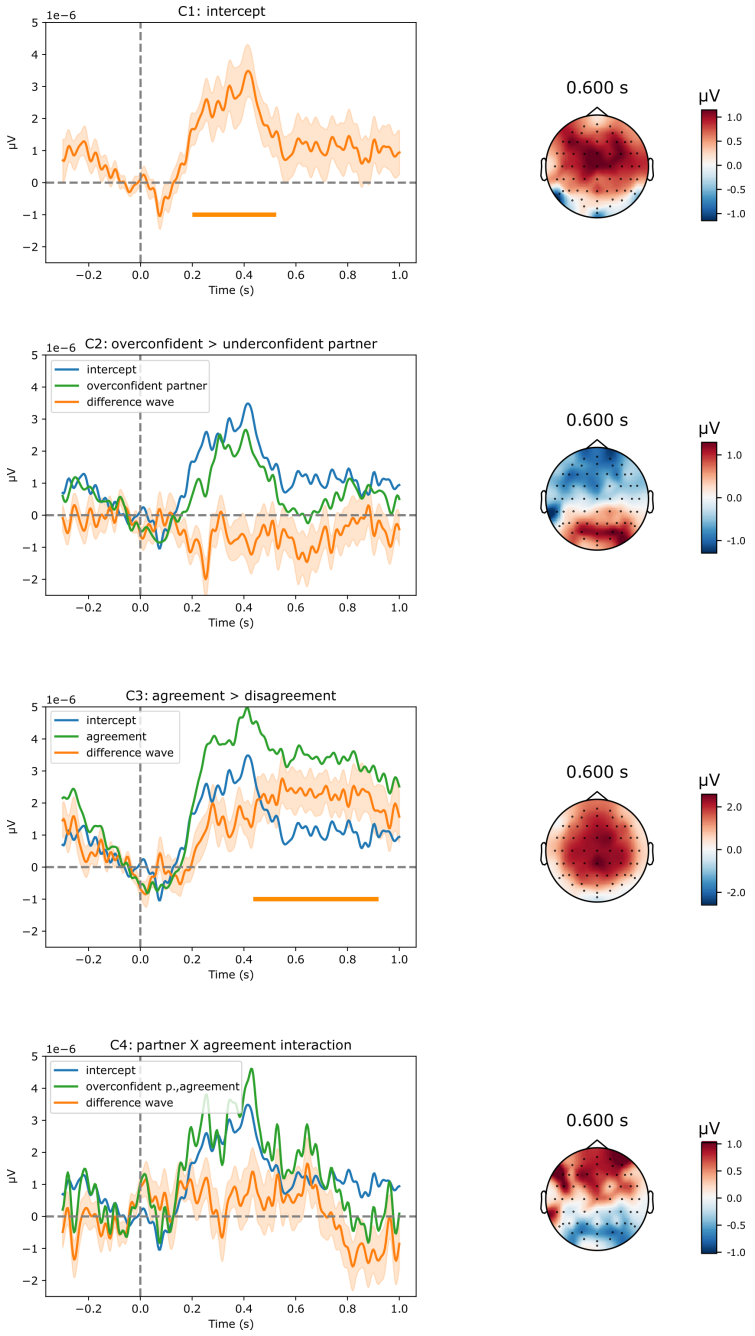


Figure 4.18: Modelled grand-average ERP at electrode FCz and visualisation of scalp topographies for the difference wave at 600 ms following stimulus presentation for the normative task condition. The ERP traces depict the mean beta coefficient across participants for the relevant condition or contrast. The intercept models the ERP for disagreement trials with the underconfident partner. Shaded areas represent the standard error of the mean. Orange bars below the x-axis denote time periods of the difference contrasts that survived permutation cluster testing against 0. C stands for contrast.

objective feedback, and a normative task, in which a joint decision was determined by the decision reported with the highest confidence, with feedback provided. Here, I aimed to use EEG to provide a continuous (rather than binary) index of participants' internal confidence, as a further test of the predicted differential impact of informational and normative influences on confidence matching.

Behaviourally, participants exhibited confidence matching in both experimental conditions. This pattern mirrors the outcomes from the experiments detailed in the preceding chapter. However, the degree of confidence matching observed in the informational task was smaller compared to the experiments in the previous chapter.

The multivariate decoding approaches employed to identify an EEG signature of confidence met with limited success. The reasons for this remain somewhat elusive. Previous studies have successfully decoded confidence on a trial-by-trial basis (e.g. Boldt and Yeung, 2015; Desender, Murphy, et al., 2019; Fernandez-Vargas et al., 2020; Kopčanová et al., 2023; Sadras et al., 2022; Steinhauser and Yeung, 2010). Moreover, a study by Steinhauser and Yeung (2010) looking at the error positivity (Pe) and its relation to error detection found similar effects for raw Pe and decoder-derived scores, but the latter were much less noisy, demonstrating that decoded values could be more precise and reliable as opposed to raw ERPs on a trial-by-trial basis. My univariate analysis revealed that the P3 component at electrode Pz correlated with confidence only when accounting for experimental conditions as covariates. In contrast, the multivariate models did not explicitly account for these factors, which could have hindered their ability to decode confidence. Given the variations in how people report their confidence, and the fact that the mapping between internal and reported confidence shifted depending on the task and partner conditions, a potential issue with the decoding approach may have been the lack of a precise true label for model training. Additionally, data was combined from two

sessions involving different perceptual decision tasks completed on separate days, introducing additional noise.

Reflecting on these findings, there are several modifications that might be implemented to the study design. Firstly, the average accuracy of participants' responses was at around 60%, which might have led to a feeling akin to guessing. This relatively low performance could have potentially contributed to the smaller confidence shifts observed between partner conditions in comparison to the previous experiments. One possibility for an improved experiment design here could be to deliberately vary the difficulty levels of the task and examine how changes in difficulty interact with partner confidence to influence participants' confidence. This modification could provide insights into the relative impact of partner confidence versus actual objective difficulty on participant confidence. Secondly, the design of the experiment could benefit from being refined to better accommodate multivariate decoding. A valuable addition could be the establishment of a baseline by training a classifier on an independent block wherein participants complete the task on their own. In this independent block, difficulty could also be deliberately varied in order to explicitly manipulate confidence. This trained classifier could subsequently be used to decode confidence during partner trials. If confidence shifts manifest predominantly at the reporting level, this approach could highlight such trends via systematic biases in classifier predictions. Thirdly, in addition to these changes, one could integrate other measures of internal confidence. For instance, incorporating techniques such as pupillometry could provide additional indices of decision confidence or cognitive effort (Lempert et al., 2015; van der Wel and Van Steenbergen, 2018). Alternatively, devising a continuous behavioural measure of internal confidence could be useful. For example, rather than relying on information seeking as in the previous experiments, participants could be requested to privately place bets on trials, providing an indirect measure of their internal confidence.

The ERP component analysis revealed a significant relationship between reported confidence and ERP amplitude at the Pz electrode when taking into account both task condition and partner condition. The late ERP amplitude (approximately 500 to 800 ms after stimulus onset) was found to be greater prior to higher reported confidence levels. This late ERP component, likely corresponding to the P3 (sometimes referred to as the centro-parietal positivity (CPP) but mostly deemed equivalent (O'connell et al., 2012)) based on its temporal and spatial profile (Kelly and O'Connell, 2013; O'connell et al., 2012; Tagliabue et al., 2019), offers further support to the hypothesis that the P3 component may function as a neural index of confidence during decision-making tasks (Fitzgerald et al., 2022; Herding et al., 2019; Rausch et al., 2020; Zakrzewski et al., 2019). A noteworthy study by Dou et al. (2023) similarly observed this correlation, even when confidence was experimentally uncoupled from accuracy, exploiting the positive evidence bias. This study posits that the P3 component signifies confidence, instead of merely denoting evidence accumulation.

Despite this association between confidence and the P3, no significant effects of task condition, partner condition, or their interaction on the P3 component were identified. Based on my initial hypotheses and consistent with the previous chapter in which information seeking indexed internal confidence, I expected a significant interaction between task and partner conditions, whereby within the informational task (but not the normative one) the P3 in the overconfident partner condition would be higher, indicating greater internal confidence, compared to the P3 in the underconfident partner condition. Additionally, when adjusting for reported confidence effects on P3 amplitude, I anticipated an interaction effect, predicting a lower P3 for the overconfident partner condition compared to the underconfident partner condition within the normative task. Although these effects were not evident, a separate GLM, which did not adjust for task and partner conditions but instead considered only

the experiment block count, did not yield a significant effect of reported confidence on ERP amplitude. This observation hints at the potential influence of partner and task conditions in imparting some meaningful variance in the association between overtly reported confidence and P3 amplitude.

One potential explanation for the absence of anticipated results is the complexity of the P3 as a marker, which may not be a definitive or exclusive indicator of internal confidence. The P3 component has been linked to numerous cognitive processes, implying that its association with confidence might be multifaceted. For instance, Salti et al. (2012) reported the P3 to correlate with conscious perception even when confidence was controlled for, and other studies have also associated the P3 with attention, working memory, or inhibitory processes (Daffner et al., 2011; Polich and Kok, 1995; Portin et al., 2000; Smith et al., 2008). It is possible that the P3 was influenced by other aspects of the task in this study. Moreover, task-dependent confidence has also been associated with other EEG markers apart from P3 amplitude. Examples include response-locked Error Positivity (Boldt and Yeung, 2015; Desender, Murphy, et al., 2019, though see Feuerriegel et al., 2022; Rausch et al., 2020), the peak latency of the stimulus-locked P3, and its slope within a window from 300 ms post-stimulus to the signal's positive peak (Shooshtari et al., 2019). Furthermore, studies have demonstrated a positive correlation between higher confidence and both spontaneous and response-locked alpha (8-12 Hz) suppression, i.e., lower alpha power (Baumgarten et al., 2016; Faivre et al., 2018; Kopčanová et al., 2023; Samaha et al., 2017).

In my exploratory analyses, I discovered an effect of partner agreement on the ERP following the partner's response at electrode FCz. Interestingly, this effect emerged not in the typical FRN window of 250 to 350 ms post feedback presentation, but rather later, approximately 500 to 900 ms post partner response presentation. Con-

sequently, it does not appear to correspond directly to the traditional FRN response. Given previous findings that have demonstrated that the FRN, typically associated with feedback processing, can also be elicited in the context of social interactions, such as when receiving counter-advice or experiencing social misalignment (Chen et al., 2012; Du et al., 2019; Shestakova et al., 2013; Tugin et al., 2015; Yu and Sun, 2013), I hypothesised that the FRN would be larger for partner disagreement when overconfident partners disagree, compared to disagreement involving underconfident partners. However, my findings do not provide evidence in support of this hypothesis. Instead, I observed a delayed effect of partner agreement on ERP amplitude with no consistent difference in this agreement effect between partner conditions. This delayed effect of agreement on ERP amplitude possibly reflects an aspect of the P3 response. However, the observed topography of this partner agreement effect, with its focus on a frontal-central location, appears to deviate from the typical parietal-central distribution of the P3. Furthermore, the fact that the ERP was larger for partner agreement compared to disagreement seems at odds with existing theories suggesting that the P3 component may reflect attention or surprise (Gaeta et al., 2003). Given that disagreement trials were less frequent, they should be more surprising and attention-demanding. This divergence might hint that the ERP response in this context is influenced by more than just surprise or attentional elements. It could be further shaped by the social complexities inherent in partner agreement dynamics. It might be that agreement, even if more common, has a particular significance or value in social contexts, which could explain its heightened neural response. The precise nature of this response and the cognitive processes it reflects, however, remain to be clarified.

In conclusion, this study expands upon the experiments in the previous chapter, providing additional support for the phenomenon of behavioural confidence matching in a social context. Moreover, the P3 ERP component emerged as a neural correlate

of confidence, further contributing to the growing body of evidence supporting this association. Notably, the relationship between confidence and P3 amplitude was only evident when taking into account task condition and partner condition, hinting at the fact that these factors may have differentially influenced overtly reported confidence versus internal confidence as reflected in the P3. Despite this, no significant effects of task and partner conditions were observed directly. Finally, it is important to consider the multifaceted nature of social influences on our beliefs and behaviours. Informational and normative influences may interact in complex and often subtle ways, especially in situations characterised by uncertainty. For instance, individuals who are unsure about the right decision or course of action might view the group's or partner's actions or beliefs not only as a valuable source of information but also as a normative standard to adhere to for fear of social rejection or misalignment.

5

Confidence Biases in Problem Gamblers

Chapter Abstract

Subjective confidence plays an important role in guiding behaviour, especially when objective feedback is unavailable. Systematic misjudgements in confidence can lead to maladaptive behaviours and have been linked to various psychiatric disorders. This study investigated confidence biases in problem gamblers compared to demographically matched control participants. Confidence was examined across different hierarchical levels of metacognition, encompassing local decision confidence, global task performance confidence, and overarching self-esteem. The problem gamblers demonstrated significantly higher local trial and global task confidence compared to control participants, despite lower self-esteem levels and after controlling for objective task performance. This overconfidence bias persisted even after controlling for the transdiagnostic symptom dimensions Anxiety-Depression and Compulsive Behaviour and Intrusive Thought, on which problem gamblers scored higher compared to control participants. The findings suggest a contrast in problem gamblers

between elevated confidence in individual decisions and overall lowered self-esteem. Additionally, the findings indicate that these features cannot be solely attributed to increased Compulsive Behaviour and Intrusive Thought and Anxiety-Depression levels. Factors such as diminished sensitivity to objective evidence, cognitive distortions, and cognitive inflexibility in problem gamblers might fuel overconfidence, thereby triggering the cycle of escalating gambling behaviours.

5.1 Introduction

One fundamental aspect of decision-making involves evaluating the utility of our choices, often in the absence of immediate external feedback. In such situations, we typically depend on an internally generated sense of confidence to guide our actions (Desender et al., 2018; Folke et al., 2016). However, a distorted sense of confidence can be problematic. If our self-perception does not align with our actual performance, it may lead us to persevere with damaging choices or behaviours, thereby undermining our ability to adapt effectively. In extreme cases, inaccurate confidence judgements that do not mirror reality can be linked to pathological behaviours. This can, for instance, manifest as extensive checking due to diminished confidence in memory (Boschen and Vuksanovic, 2007; Tolin et al., 2001), or delusional thinking stemming from overconfidence in false beliefs (Moritz et al., 2005). Hence, a thorough understanding of metacognition — the monitoring and control of our own thoughts and behaviours — and its deficiencies plays a crucial role in the study and treatment of various psychiatric disorders (see Chapter 1; David et al., 2014; Sharma et al., 2022; Sun et al., 2017).

Case-control studies have uncovered distinct patterns of alterations in metacognition across several psychiatric disorders. For example, alterations have been observed in depression (Fieker et al., 2016; Fu et al., 2005; Hancock, 1996), obsessive-compulsive disorder (OCD; Macdonald et al., 1997; McNally and Kohlbeck, 1993; Moritz and Jaeger, 2018), and psychosis (Eifler et al., 2015; Li et al., 2018; Lysaker et al., 2020; Moritz et al., 2005; for a review, see Hoven et al., 2023). However, given the high heterogeneity within and comorbidity between disorders, the nascent field of transdiagnostic psychiatry proposes that cognitive, affective, and neurobiological processes that govern complex behaviour may correspond more closely with transdiagnostic symptom dimensions rather than conventional diagnostic categories (Wise et al.,

2023). This approach transcends traditional diagnostic classifications, such as those found in the Diagnostic and Statistical Manual of Mental Disorders (DSM; American Psychiatric Association, 2022), and potentially offers a more nuanced understanding of psychiatric conditions (Huys et al., 2016; Stephan and Mathys, 2014).

A study by Rouault, Seow, et al. (2018) leveraged this transdiagnostic approach to investigate the relationship between confidence and psychiatric symptomatology in a non-clinical sample. Participants performed a perceptual decision-making task and completed a series of self-report questionnaires capturing a broad spectrum of psychiatric symptoms, including depression, general anxiety, schizotypy, impulsivity, OCD, social anxiety, eating disorders, apathy, and alcohol dependency. The researchers performed a factor analysis to determine a concise latent structure that best explained the variance at the questionnaire item level, leading to the identification of three symptom dimensions: Anxiety-Depression, Compulsive Behaviour and Intrusive Thought, and Social Withdrawal. These dimensions were consistent with the factor structure first identified by Gillan et al. (2016) using the same set of psychiatric symptom questionnaires. The Anxiety-Depression dimension, which primarily links apathy, anxiety, and depression features, showed a significant correlation with lower mean confidence and higher metacognitive efficiency (the ability to distinguish between one's own correct and incorrect judgements given a certain level of task performance). In contrast, the Compulsive Behaviour and Intrusive Thought dimension, characterised predominantly by elements of impulsivity, OCD, schizotypy, addiction, and eating disorders, was linked with higher mean confidence and lower metacognitive efficiency.

These diametrically opposed associations between the Anxiety-Depression and Compulsive Behaviour and Intrusive Thought dimensions with confidence – whereby Anxiety-Depression is linked to decreased confidence and Compulsive Behaviour

and Intrusive Thought to increased confidence – have been replicated in numerous studies (Benwell et al., 2022; Fox et al., 2023; Hoven et al., 2023; Rouault, Seow, et al., 2018; Seow and Gillan, 2020). Importantly, these associations do not emerge when only looking at individual questionnaire data, underscoring the merit of employing a transdiagnostic framework to account for existing comorbidities in psychopathologies. For instance, patients diagnosed with OCD may not only exhibit symptoms of compulsivity but also anxiety to a greater or lesser extent (Bartz and Hollander, 2006; Tükel et al., 2002), which could obscure findings related to the underlying cognitive processes when considering solely the traditional DSM disorder categories. This was also demonstrated in a study by Gillan et al. (2020), finding that the Compulsive Behaviour and Intrusive Thought dimension was a significant predictor of deficits in goal-directed planning, whereas a diagnosis of OCD alone was not.

The seemingly paradoxical phenomenon, whereby the symptom dimensions of Anxiety-Depression and Compulsive Behaviour and Intrusive Thought demonstrate opposing associations with confidence yet are positively correlated, suggests that confidence is likely influenced by multiple cognitive processes. In order to unpack this idea, it is useful to consider multiple hierarchical levels at which metacognition operates, spanning from confidence in individual decisions (local confidence), to confidence in overall task performance (global confidence), to even higher-order attributes such as self-esteem (Seow et al., 2021). Indeed, it has been suggested that the confidence biases may be driven by two distinct mechanisms; reduced confidence related to Anxiety-Depression may originate from global self-beliefs like self-esteem, whereas overconfidence related to Compulsive Behaviour and Intrusive Thought could be a consequence of difficulties in constructing an accurate mental model of one's own performance (Seow et al., 2021). Supporting this idea, Hoven et al. (2023) found a negative association between Anxiety-Depression and self-esteem but not local con-

fidence, whereas Compulsive Behaviour and Intrusive Thought was positively linked with local confidence but negatively with self-esteem.

Within this framework, more globally-held beliefs are formed over extended periods of time, integrating information gleaned from a multitude of lower-level experiences. Consequently, metacognitive information obtained from monitoring lower-level activities can be employed as feedback to infer higher-order self-beliefs, such as self-efficacy and mastery, which may in turn influence our approach to life and our perceived control over it (Bandura, 1978; O'Kearney et al., 2020). These beliefs are important determinants of mental health (Quiles et al., 2015; Silverstone and Salsali, 2003). Conversely, the prevailing set of more globally held self-beliefs may also shape an individual's local confidence in specific decisions. This process of monitoring our actions and cognitive processes on multiple levels may also manifest itself in metacognitive control operating at different levels and on different time-scales. For instance, at the local level, local confidence may inform information seeking choices for individual decisions (Desender et al., 2018), whereas at the global level, global confidence may guide the selection of tasks to perform (Carlebach and Yeung, 2020; Rouault et al., 2019). These examples underscore the dynamic interplay between different levels of metacognition in shaping our beliefs and behaviour (Hoven et al., 2023).

The current study aimed to probe the nature and extent of maladaptive metacognition within a sample of problem gamblers. Gambling disorder is a condition currently classified as a substance-related and addictive disorder by the DSM-5 (American Psychiatric Association, Association, et al., 2013). This classification highlights the similarities between gambling disorder and substance use disorders, with shared features that include chronic progression, high rate of relapse, and fundamental changes in the brain's reward and motivational systems (Goudriaan et al., 2019).

Gambling disorder is characterised by persistent and recurrent problematic gambling behaviour leading to substantial impairment or distress. Of specific interest in this study was whether the double-dissociation observed in the general population, wherein the Anxiety-Depression dimension predicts underconfidence and the Compulsive Behaviour and Intrusive Thought dimension predicts overconfidence, is also prevalent among problem gamblers. Furthermore, I aimed to examine how these associations may manifest at different levels of the metacognitive hierarchy.

Problem gamblers present a particularly interesting case study given that problem gambling is typified by compulsive behaviour, yet also demonstrates high comorbidity rates with anxiety and depression (Dannon et al., 2006; Dowling et al., 2015; Jauregui et al., 2016; Lorains et al., 2011). Building on the findings described above, I hypothesised that Anxiety-Depression would correlate with reduced confidence at the self-esteem level, whereas Compulsive Behaviour and Intrusive Thought would be associated with elevated confidence at the local decision level, such that higher symptom severity might trigger a dissociation of the different levels of the metacognitive hierarchy. Previous investigations into confidence levels in problem gamblers have pointed towards a tendency for overconfidence among these individuals (Brevers et al., 2014; Goodie, 2005; Lakey et al., 2007). However, these studies failed to adequately control for performance differences, thus rendering definitive conclusions about confidence biases difficult to substantiate. In a recent study by Hoven, de Boer, et al. (2022), it was reported that there was no significant difference in confidence levels between patients with gambling disorder and a healthy control sample, after attempting to control for gender. This control is important given that gambling disorder predominantly affects males (Howe et al., 2019; Subramaniam et al., 2015; Welte et al., 2017), and males have been observed to generally exhibit higher levels of confidence (Ariel et al., 2018; Rivers et al., 2021). However, it is crucial to note that an inherent challenge existed in truly controlling for gender due to it being

severely confounded with the clinical status, whereby the gambling disorder group was overwhelmingly composed of males, in contrast to the more gender-balanced control group. In light of these complexities, comparing problem gamblers to a demographically matched group of control participants and adopting a transdiagnostic approach to examine variations in symptoms could offer more robust insights into potential abnormalities at different levels of the metacognitive hierarchy associated with problem gambling.

5.2 Methods

Participants. After applying all exclusion criteria, the sample consisted of 38 problem gamblers and 38 control participants. The exclusion criteria were delineated as follows: Participants were excluded if, in the metacognition task, they consistently opted for the same side (either left or right) in over 95% of trials; if their average accuracy in the metacognition task either fell below 60% or exceeded 85%; if there were disparities in the data provided for participants' gender, country of residence, or age within a range of $\pm$ 1 year between the preliminary screening survey and the data compiled from the Neureka app in which participants completed the metacognition task; if the participant was not residing in the United Kingdom, Ireland, or the United States; or finally, if participants, being potential controls, were females under the age of 21. This last criterion was adopted to ensure a demographic match between control participants and problem gamblers.

Table 5.1 depicts the demographic and psychological measures of problem gamblers and control participants as well as between-groups comparisons. Participants were compensated with a €10 gift card upon full completion of the task. Recruitment was done via online forums, posters displayed near gambling venues, university mailing lists, and word of mouth. The study was approved by the School of Psychology

Research Ethics Committee, Trinity College Dublin.

Characteristic	PG	CP	$\chi^2/t(df)$	p
Gender, $n(\%)$			2.19 (1)	.139
Male	34 (89.5)	28 (73.7)		
Female	4 (10.5)	10 (26.3)		
Country of residence, $n(\%)$			2.83 (2)	.243
Ireland	18 (47.4)	25 (65.8)		
United Kingdom	2 (5.3)	2 (5.3)		
United States	18 (47.4)	11 (28.9)		
Highest education level, $n(\%)$			2.33 (2)	.312
Secondary school	3 (7.9)	2 (5.3)		
University degree or equiv.	33 (86.8)	36 (94.7)		
PhD or equiv.	2 (5.3)	0 (0.0)		
Age, $M(SD)$	31.2 (6.9)	28.8 (10.7)	1.15 (63.1)	.256
RSES, $M(SD)$	14.50 (4.6)	16.79 (6.6)	1.75 (66.6)	.084
AD, $M(SD)$	0.09 (0.9)	-0.37 (0.9)	2.26 (73.9)	$< .05$
CIT, $M(SD)$	0.92 (0.9)	0.04 (0.9)	4.42 (74.0)	$< .0001$
PGSI, $M(SD)$	15.7 (4.8)	0.0 (0.0)	20.22 (37.0)	$< .0001$

Table 5.1: Demographic (gender, country of residence, highest education level, age) and psychological (Rosenberg Self-Esteem Scale (RSES), Anxiety-Depression (AD), Compulsive Behaviour and Intrusive Thought (CIT), Problem Gambling Severity Index (PGSI) measures for problem gamblers (PG) and control participants (CP).

Procedure. *Problem Gambling Severity Index.* Study volunteers were initially screened via the Problem Gambling Severity Index (PGSI) to be included as problem gamblers (PGSI $\geq$ 8) or control participants (PGSI = 0). The PGSI, a 9-item refined version of the Canadian Problem Gambling Index (CPGI; Ferris and Wynne, 2001), is a non-clinical assessment survey for problem gambling and has been used worldwide in population-level survey research (Dune et al., 2017; Kavli and Berntsen,

2005; Kristiansen and Jensen, 2011; Meyer et al., 2009; Volberg and Bernhard, 2006; Wardle, 2007; Young and Stevens, 2008). The survey asks participants to self-assess their gambling behaviour over the past 12 months by rating their agreement with statements such as '*Have you borrowed money or sold anything to gamble?*' or '*Have you needed to gamble with larger amounts of money to get the same feeling of excitement?*'. The PGSI employs a scoring system to categorise gamblers based on their behaviour and the consequences of their gambling. A score of zero is assigned to non-gamblers or those who gamble without negative consequences. Scores of 1-2 represent individuals who encounter a low level of problems with no or only minimal negative repercussions. Those scoring between 3 and 7 are considered to be experiencing a moderate level of problems, which are associated with certain negative consequences. Finally, individuals scoring 8 or above are typically facing severe gambling issues, characterised by adverse outcomes and a potential loss of control.

Rosenberg Self-Esteem Scale. Following the PGSI, participants were asked to complete the Rosenberg Self-Esteem Scale (RSES; Rosenberg, 1965). The RSES is a widely used instrument designed to measure self-esteem, consisting of ten statements related to overall feelings of self-worth or self-acceptance. The statements are designed to be answered using a 4-point Likert scale ranging from '*strongly agree*' to '*strongly disagree*'. Half of these statements have positively worded propositions (e.g., '*I feel that I'm a person of worth, at least on an equal plane with others.*'), whereas the other half contain negatively worded ones (e.g., '*I feel I do not have much to be proud of.*'). The scores from these ten items are summed up to form a total self-esteem score which can range from 0 to 30. Higher scores indicate higher self-esteem, whereas lower scores suggest lower self-esteem.

Metacognition Task. After completing the PGSI and the RSES, participants were

asked to download the Neureka app (https://www.neureka.ie/). The Neureka app, developed by the Gillan Lab at Trinity College Dublin, features a collection of gamified versions of commonly-used psychological tasks and questionnaires. Within this app, participants were asked to complete the Metamind task (see Figure 5.1). The Metamind task is a gamified version of the Dot Discrimination Task, a perceptual decision-task frequently used to measure metacognition (see Chapters 3 and 4; cf. Boldt and Yeung, 2015; Rouault, Seow, et al., 2018). In comparing Metamind to the traditional dots-task, it has been shown to have adequate validity and excellent reliability (Fox et al., 2022). In Metamind, participants are given the task of controlling a spaceship traversing through space. Upon the appearance of two objects, the challenge is to navigate the spaceship to the object containing more dots. Participants make their selection by tapping either the left or right side of their smartphone screen, corresponding to their chosen object. Following this, participants indicate how confident they are in the accuracy of their choice on a 6-point scale. Following 20 practice trials, participants perform 80 trials divided into four blocks. After every block of 20 trials, participants are asked to report their confidence in their performance in that block on a 6-point scale. Task performance is kept at ca. 72% accuracy by using a two-down-one-up log-adaptive staircasing procedure, whereby the difference in the number of dots increases (the task becomes easier) after an incorrect response and decreases (the task becomes more difficult) after two consecutive correct responses. In this task, metacognitive bias is operationalised as mean confidence. I focus on mean confidence in this study because the quantification of metacognitive sensitivity and efficiency remains a contested question. Current measures provide sub-optimal validity and reliability, and require higher trial numbers than I had available in the metacognition task employed in this study (Arnold et al., 2023; Desender et al., 2022; Guggenmos, 2021; Rahnev, 2023).

Transdiagnostic Symptom Dimensions. Finally, participants were asked to complete

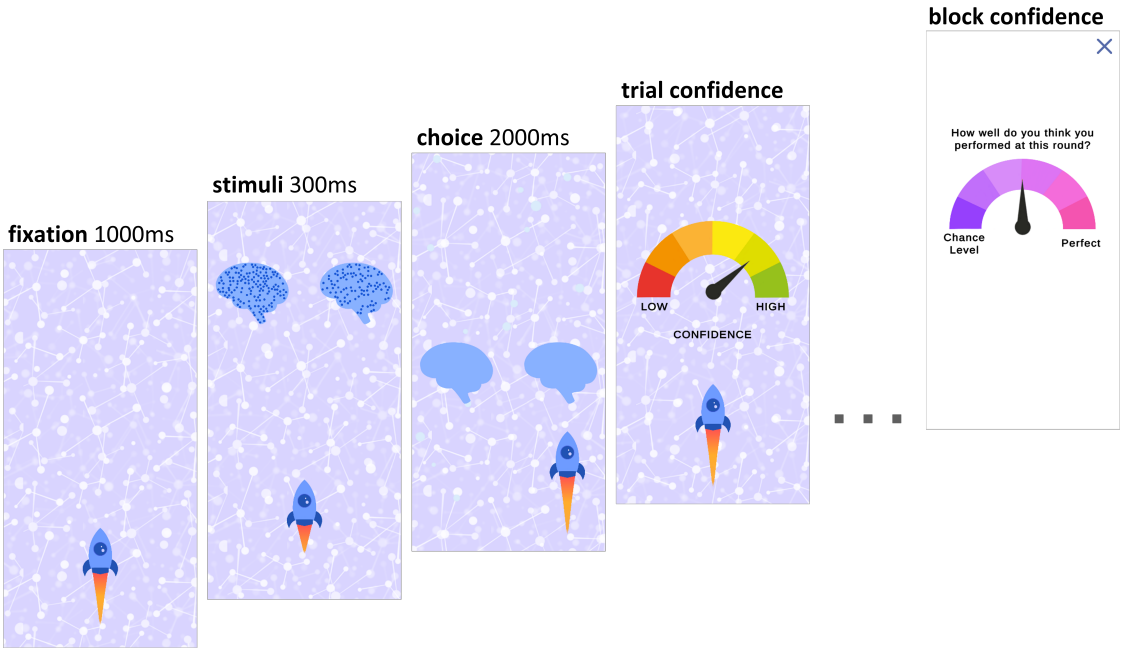


Figure 5.1: Illustration of Metamind's task structure — a smartphone game designed to evaluate metacognition. Participants were placed in control of a spaceship voyaging through space. When two objects appeared, the task was to steer the spaceship towards the object displaying a greater number of dots. This was achieved by tapping the left or right side of the smartphone screen, corresponding to the object of choice. Subsequently, participants were prompted to report their confidence in the accuracy of their choice on a 6-point scale. After the completion of every set of 20 trials, participants were further asked to report their confidence in their overall performance across the preceding block of 20 trials on a 6-point scale.

a range of psychiatric measures in order to derive Anxiety-Depression and Compulsive Behaviour and Intrusive Thought scores, two of the three transdiagnostic factors identified by Gillan et al. (2016). To measure these factors more efficiently, I used a reduced set of questions that has been shown to provide an accurate approximation of the true factor scores (Wise and Dolan, 2020). I included only those questionnaires that pertain specifically to the Anxiety-Depression and Compulsive Behaviour and Intrusive Thought dimensions. Accordingly, participants completed the following questionnaires: Apathy Evaluation Scale (AES, Marin et al., 1991), Barrett's Impulsivity Scale (BIS Patton et al., 1995), Eating Attitudes Test (EAT, Garner et al., 1982), Obsessive Compulsive Inventory (OCI, Foa et al., 2002), Self-rating Depression Scale (SDS, Zung, 1965), and State Trait Anxiety Inventory (STAI, Speilberger et al., 1983). Anxiety-Depression and Compulsive Behaviour and Intrusive Thought scores were derived by using the factor weights as per Wise and Dolan (2020). Anxiety-Depression and Compulsive Behaviour and Intrusive Thought scores are scaled around 0, with higher scores corresponding to higher symptom levels.

To provide insight into what the Anxiety-Depression factor assesses, consider the three highest scoring items from this dimension: The first is derived from the AES, which inquires about participants' thoughts, emotions, and activities over the preceding four weeks. The statement is '*I have motivation*' and is coded in reverse. The second item stems from the SDS, requesting participants to express how they felt or behaved in the past few days. The statement is '*I feel that I am useful and needed*' and it is also reverse-coded. The third item is extracted from the STAI, probing into how participants generally feel. The statement is '*I feel satisfied with myself*' and is coded in reverse as well. For the Compulsive Behaviour and Intrusive Thought dimension, the three highest scoring items are as follows: The first two are from the OCI, which asks participants how much they have been distressed or bothered by a particular experience in the previous month. The statements are '*I*

find it difficult to control my own thoughts' and '*I am upset by unpleasant thoughts that come into my mind against my will*'. The third item comes from the EAT and reads, '*I am terrified about being overweight*'.

5.3 Results

The characteristics of the participants are presented Table 5.1. There were no significant differences between the groups regarding the distribution of gender, country of residence, level of education, or age. This suggests successful matching of the control participants and problem gamblers with regards to demographic variables. The problem gamblers exhibited significantly higher Anxiety-Depression and Compulsive Behaviour and Intrusive Thought scores compared to the control participants (see Figure 5.2). Within the problem gamblers, problem gambling severity was positively, albeit non-significantly, associated with Compulsive Behaviour and Intrusive Thought symptoms ($r = 0.21, p = .213$), but showed no association with Anxiety-Depression symptom levels ($r = 0.00, p = .996$; Figure 5.3). Despite the staircasing procedure, control participants performed at a slightly higher mean accuracy of 73.2% (±0.03, range = 68.8 — 77.5) compared to problem gamblers at 71.7% (±0.04, range = 62.5 — 77.5; $t(74) = 2.1, p < .05$). To account for potential effects of task accuracy as well as age and gender, these parameters are included as covariates in subsequent analyses where relevant. For a comprehensive understanding of all descriptive and between-group comparison statistics, refer to Table 5.1.

My key question was whether problem gamblers would show differences in confidence bias relative to controls. To address this question, linear regression analyses were performed to test for group differences in local confidence and global confidence, while controlling for gender, age and mean task accuracy, as well as for self-esteem,

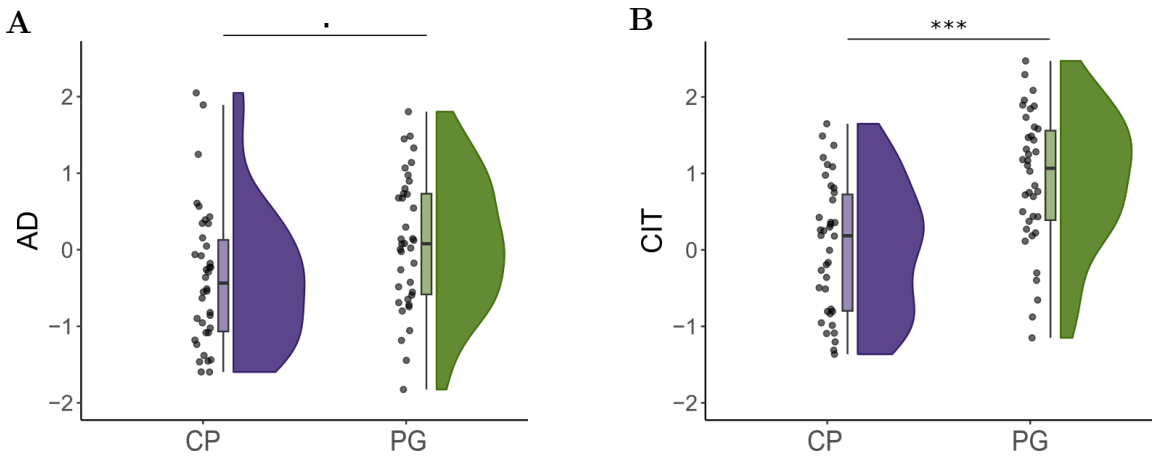


Figure 5.2: **A.** Anxiety-Depression (AD) and **B.** Compulsive Behaviour and Intrusive Thought (CIT) scores for control participants (CP) and problem gamblers (PG). Dots show data from individual participants. Violin and box plots show the distributions of participant means. · $p < .05$, *$p < .01$, **$p < .001$, ***$p < .0001$ in two-sample T-test.

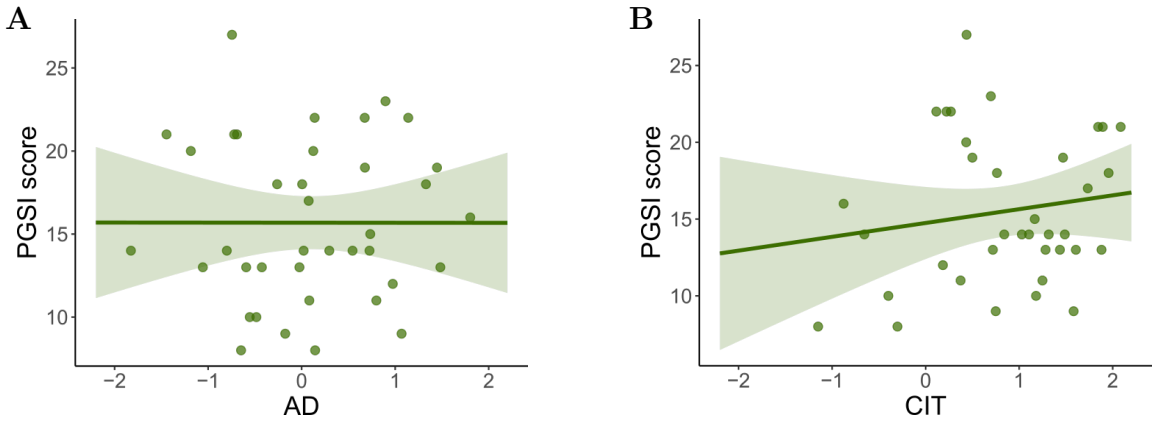


Figure 5.3: Relationship of **A.** Anxiety-Depression (AD) and **B.** Compulsive Behaviour and Intrusive Thought (CIT) with Problem Gambling Severity Index (PGSI) score in problem gamblers

while controlling for age and gender. These analyses revealed significant effects of group on local confidence ($\beta = 0.91, SE = 0.20, p < .0001$), whereby problem gamblers reported significantly higher confidence at the trial-level compared to control participants, and on global confidence ($\beta = 1.08, SE = 0.28, p < .001$), whereby problem gamblers reported significantly higher confidence at the block-level compared to control participants (see Figure 5.4 A & B). There were no significant effects of gender, age or mean task accuracy on local and global confidence (all $p > .27$). The higher confidence was not reflected in any significant difference between mean reaction times between problem gamblers ($M = 0.91$ seconds) and control participants ($M = 0.94$ seconds; $t(74) = 0.90, p = .373$; Figure 5.4 D). In contrast to local and global confidence, self-esteem was lower for problem gamblers compared to control participants. However, this group effect on self-esteem was not significant when controlling for gender and age ($\beta = -2.4, SE = 1.33, p = .072$, Figure 5.4 C). There were no significant effects of gender or age on self-esteem (both $p > .12$).

Next, I examined whether the observed group differences in confidence bias could be explained by the transdiagnostic symptom dimensions Compulsive Behaviour and Intrusive Thought and Anxiety-Depression (see Figure 5.5). Specifically, I predicted that Compulsive Behaviour and Intrusive Thought would correlate with elevated confidence, potentially explaining the higher confidence at the local trial and global task level in problem gamblers compared to control participants. Additionally, I predicted that Anxiety-Depression would correlate with reduced confidence, potentially accounting for lower self-esteem. Following the format of my previous analyses, I constructed three regression models to predict local confidence, global confidence, and self-esteem. Instead of group, Anxiety-Depression and Compulsive Behaviour and Intrusive Thought scores were used as predictors. As expected, there was a significant positive association between Compulsive Behaviour and Intrusive

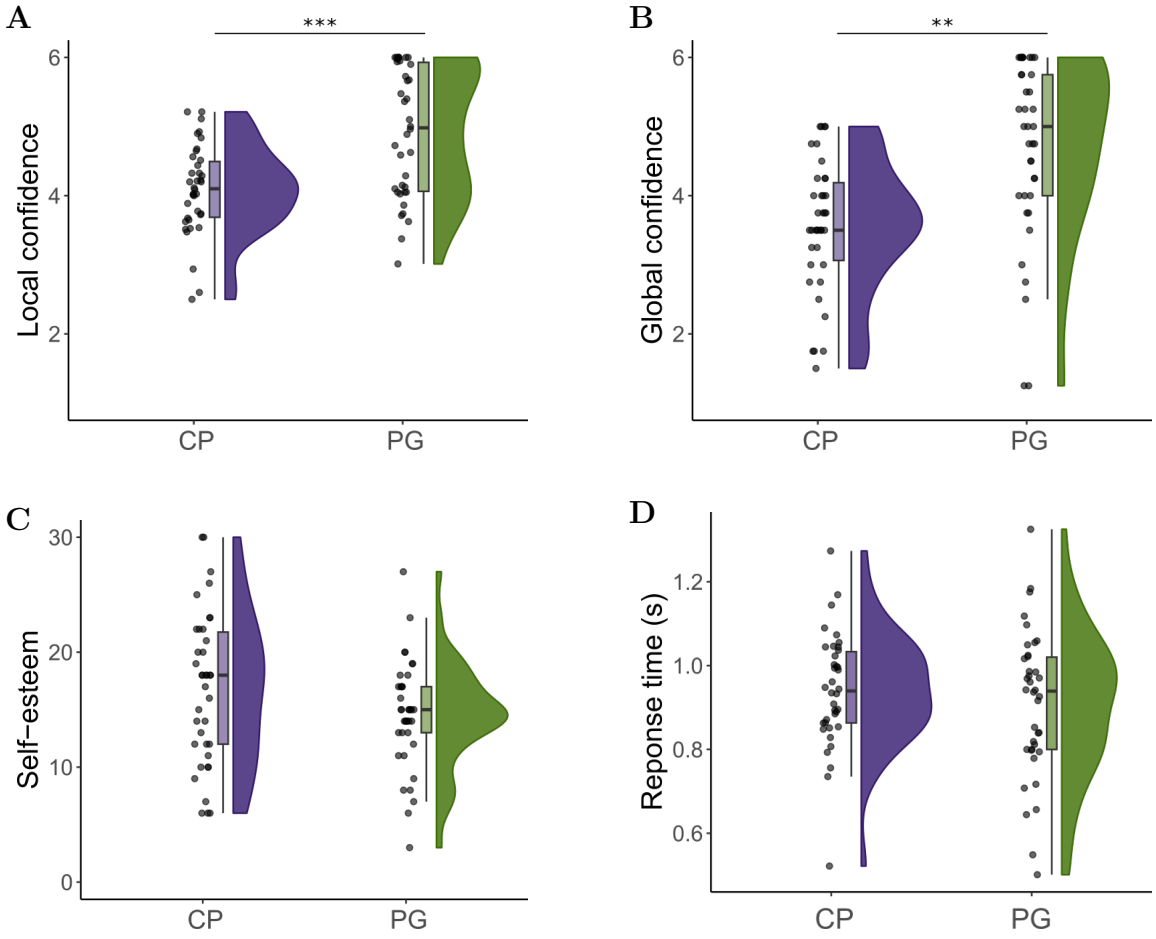


Figure 5.4: **A.** Local trial confidence, **B.** global task confidence, and **C.** self-esteem for control participants (CP) and problem gamblers (PG). Dots show data from individual participants. Violin and box plots show the distributions of participant means. $^{*}p < .01$, $^{**}p < .001$, $^{***}p < .0001$ in linear regression with age, gender, and task accuracy as covariates. **D.** Mean reaction times in seconds for control participants (CP) and problem gamblers (PG)

Thought and local confidence ($\beta = 0.24, SE = 0.11, p < .05$), but no effect of Anxiety-Depression on local confidence ($\beta = 0.01, SE = 0.11, p = .928$). These results were mirrored in the regression on global confidence, with a significant positive association between global confidence and Compulsive Behaviour and Intrusive Thought ($\beta = 0.34, SE = 0.15, p < .05$), but no effect of Anxiety-Depression on global confidence ($\beta = 0.02, SE = 0.15, p = .917$). Contrary to local and global confidence, variations in self-esteem were not significantly explained by either Compulsive Behaviour and Intrusive Thought ($\beta = -0.34, SE = 0.70, p = .628$) or Anxiety-Depression ($\beta = -0.66, SE = 0.69, p = .341$). None of the covariates (gender, age, and task accuracy) was statistically significant in any of the three regression models (all $p > .076$).

Building on these findings, I again constructed three separate regression models to predict each of local confidence, global confidence, and self-esteem. This time, I included both group and the transdiagnostic symptom dimensions, Anxiety-Depression and Compulsive Behaviour and Intrusive Thought, as predictors to establish which among these was best at explaining the observed differences in confidence bias. Notably, the significant group effects on local confidence ($\beta = 0.91, SE = 0.23, p < .001$) and global confidence ($\beta = 1.02, SE = 0.32, p < .005$) remained despite including Anxiety-Depression and Compulsive Behaviour and Intrusive Thought as predictors in the regression models. Conversely, there was no longer a significant effect of Compulsive Behaviour and Intrusive Thought on local ($\beta = 0.05, SE = 0.11, p = .654$) and global ($\beta = 0.12, SE = 0.16, p = .441$) confidence. As before, the effects of Anxiety-Depression on local ($\beta = -0.08, SE = 0.10, p = .408$) and global ($\beta = -0.09, SE = 0.14, p = .534$) confidence were not significant in the regression models. In the case of self-esteem, there were no significant effects of any of group ($\beta = -2.4, SE = 1.58, p = .135$), Compulsive Behaviour and Intrusive Thought ($\beta = 0.21, SE = 0.78, p = .791$) or Anxiety-Depression

($\beta = -0.44, SE = 0.70, p = .533$). None of the covariates (gender, age, and task accuracy) were significant in any of the three regression models (all $p > .15$).

Lastly, I investigated the inter-relationship between local trial confidence, global task confidence, and self-esteem, with a particular focus on potential differences in these relationships between problem gamblers and control participants (see Figure 5.6). Both the problem gamblers and control participants exhibited a strong association between local trial confidence and global task confidence (problem gamblers: $r = 0.71, p < .0001$; control participants: $r = 0.75, p < .0001$). Applying Fisher's Z-test to examine whether the strength of these correlations differed between the two groups showed that the correlation coefficients did not significantly differ ($Z = -0.33, p = .74$). In contrast, self-esteem appeared to be largely independent of local confidence in the control participants ($r = -0.08, p = .629$), but was positively associated with local confidence in the problem gamblers ($r = 0.33, p < .05$; $z = 1.78, p = .075$). Self-esteem was not significantly correlated with global confidence in either group (problem gamblers: $r = 0.19, p = .251$; control participants: $r = -0.11, p = .522$; $Z = 1.26, p = .208$).

5.4 Discussion

Our internal sense of confidence plays a crucial role in guiding our behaviours, particularly when external feedback is lacking. Misjudgements in confidence can result in maladaptive behaviours, and systematic aberrations have been associated with psychiatric disorders. The transdiagnostic approach, which characterises symptoms across diagnostic boundaries rather than adhering to traditional diagnostic categories, has shown that a symptom dimension of Anxiety-Depression is associated with reduced confidence, whereas a Compulsive Behaviour and Intrusive Thought dimension is associated with elevated confidence (Rouault, Seow, et al., 2018). This

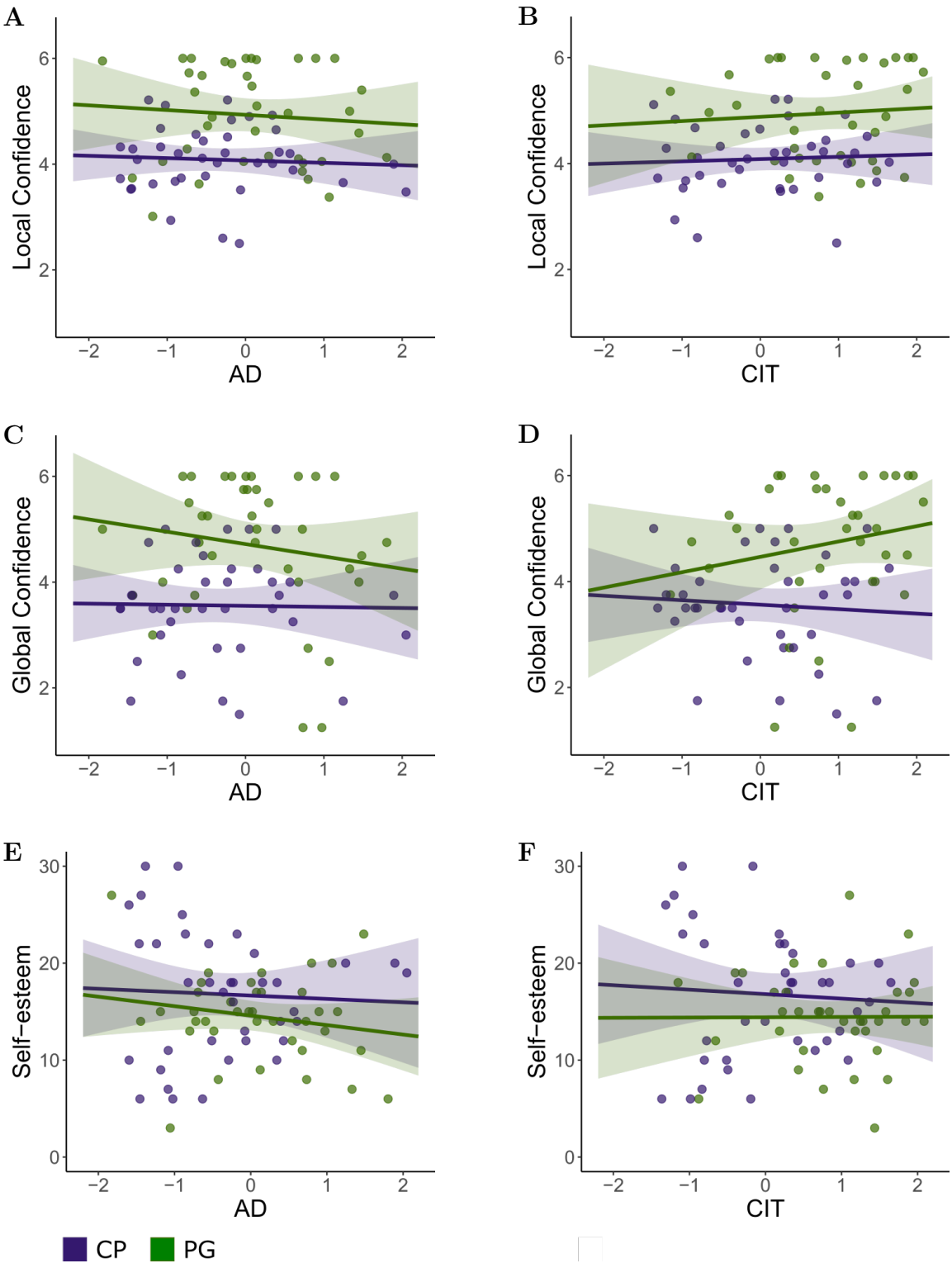


Figure 5.5: Relationship between the transdiagnostic symptom dimensions (Anxiety-Depression (AD) and Compulsive Behaviour and Intrusive Thought (CIT) and confidence at all levels of the metacognitive hierarchy (local confidence, global confidence, and self-esteem) for control participants (CP) and problem gamblers (PG).

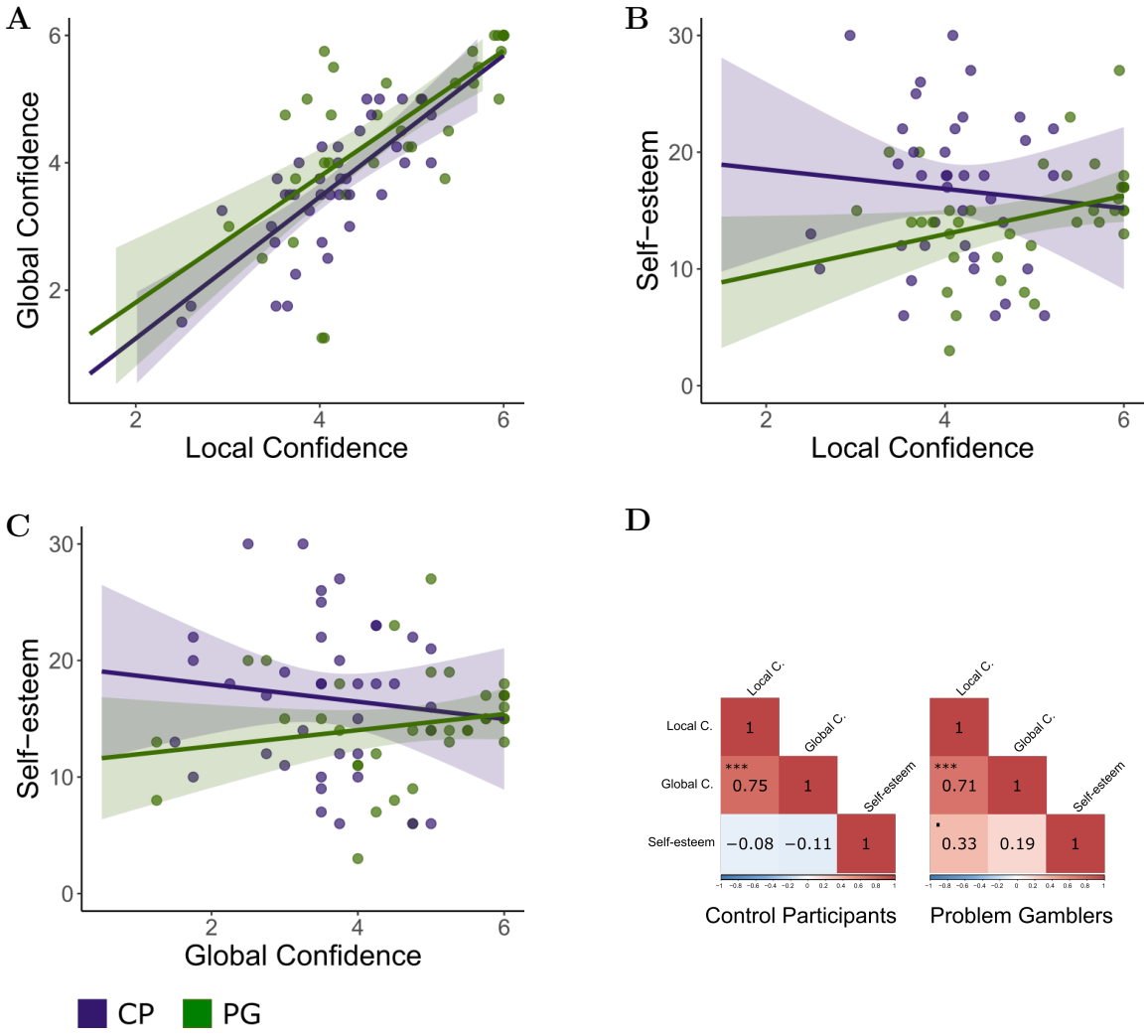


Figure 5.6: **A.-C.** Relationship between local trial confidence, global task confidence, and self-esteem for control participants (CP) and problem gamblers (PG). **D.** Correlation matrix for local trial confidence, global task confidence, and self-esteem for control participants and problem gamblers. $\cdot\ p < .05$, $^{*}p < .01$, $^{**}p < .001$, $^{***}p < .0001$

study sought to investigate how these confidence alterations manifest at different hierarchical levels of metacognition (local confidence, global confidence, self-esteem) in problem gamblers, a group often displaying symptoms of both Anxiety-Depression and Compulsive Behaviour and Intrusive Thought, compared to control participants.

The findings demonstrated that a group of problem gamblers showed significantly higher local trial-by-trial and global task confidence compared to control participants, even after controlling for gender, age, and objective task accuracy. However, despite the problem gamblers' elevated confidence on trial and task levels, their overall self-esteem was generally lower than that of the control participant group (albeit not significantly so). I hypothesised that the heightened confidence within the problem gamblers might be attributable to a positive correlation with elevated Compulsive Behaviour and Intrusive Thought levels in problem gamblers, whereas the diminished self-esteem might be associated with increased levels of Anxiety-Depression in this group. Although I observed the expected significant association of Compulsive Behaviour and Intrusive Thought with elevated local and global confidence across groups, this effect diminished when controlling for group. Moreover, there was no significant effect of Anxiety-Depression on confidence at any level of the metacognitive hierarchy. The group effect on elevated confidence on the other hand remained significant even when controlling for the transdiagnostic symptom dimensions, Anxiety-Depression and Compulsive Behaviour and Intrusive Thought. This suggests that there are differences between the problem gamblers and control participants driving elevated decision confidence that are not captured by the transdiagnostic symptom dimensions.

The observation of significantly higher levels of local and global confidence in the problem gamblers in comparison to the control participants, even after accounting for elevated levels of Compulsive Behaviour and Intrusive Thought and Anxiety-

Depression, and despite lower self-esteem, raises intriguing questions about the underlying mechanisms contributing to heightened decision confidence in this group. Research conducted by Hoven, de Boer, et al. (2022) found that problem gamblers displayed a reduced integration of evidence into their confidence judgements for correct choices. This was observed when compared to both healthy controls and OCD patients, a comparison that underlines the presence of additional processes specific to problem gamblers, given that OCD patients also display high Compulsive Behaviour and Intrusive Thought symptom levels. This diminished sensitivity towards objective evidence might align with cognitive distortions that are a common occurrence in problem gamblers. These distortions may include biases like interpretive bias (perceived ability to interpret or control ambiguous events), illusion of control (overestimation of ability to control events), or predictive control (reflecting probability errors such as the gamblers' fallacy; Cowley et al., 2015; Goodie, 2005; Goodie and Fortune, 2013; Ledgerwood et al., 2020; Mallorquí-Bagué et al., 2019; Orgaz et al., 2013). Moreover, problem gamblers often display cognitive inflexibility, which may include a reduced capacity to shift attention and could make them less receptive to objective evidence that contradicts their beliefs, thereby fostering overconfidence (Perandrés-Gómez et al., 2021; Van Timmeren et al., 2018). Possibly supporting the notion of a lack of sensitivity to belief-contradicting evidence, a study by Wyckmans et al. (2019) found that individuals with problem gambling disorder demonstrated impaired model-based learning, especially after non-rewarded outcomes. These individuals also exhibited faster reaction times compared to control participants following non-rewarded decisions. This lack of reduced speed in response after a loss in problem gamblers was also observed by Goudriaan et al. (2005). Such behaviour has also been associated with increased impulsive responding often observed in problem gamblers (Verdejo-García et al., 2008). However, results in the current study did not reveal any differences in mean reaction times

between the problem gamblers and control participants. This lack of a reaction time difference suggests that impulsivity, as measured by response times, may not have been a direct contributor to the observed overconfidence in problem gamblers in the current study.

Although group effects persisted even after accounting for Anxiety-Depression and Compulsive Behaviour and Intrusive Thought, these effects were smaller than those found when not accounting for the transdiagnostic dimensions. This finding indicates that, although the differences in confidence levels and self-esteem between problem gamblers and control participants are not exhaustively captured by the Anxiety-Depression and Compulsive Behaviour and Intrusive Thought symptom dimensions, these factors do explain some of the observed variance. Moreover, a regression model not including group as a predictor showed significant effects of Compulsive Behaviour and Intrusive Thought on elevated local and global confidence. Overconfidence linked to Compulsive Behaviour and Intrusive Thought has been suggested to reflect difficulties in developing an accurate cognitive map or model of the task environment (Gillan et al., 2016). Evidence for this comes from Seow and Gillan (2020), who demonstrated that individuals with higher Compulsive Behaviour and Intrusive Thought were less likely to use evidence to inform their confidence evaluations, exhibiting overall inflated confidence estimates and an inability to adequately utilise unexpected outcomes, belief uncertainty, and positive feedback to appropriately inform their confidence levels. This begs the question, if environmental evidence is not informing confidence in those high in Compulsive Behaviour and Intrusive Thought, what is? One speculative answer to this question may lie in an individual's prior expectations. Individuals with higher Compulsive Behaviour and Intrusive Thought symptoms could be basing their confidence on a distorted prior expectation of success, and thus not adequately use objective evidence available in the task environment to update their beliefs.

The lack of a clear effect of Anxiety-Depression on confidence may need to be considered in light of the effect sizes of the associations between Anxiety-Depression and reduced confidence, and Compulsive Behaviour and Intrusive Thought and elevated confidence reported in previous studies (Rouault, Seow, et al., 2018; Seow and Gillan, 2020). Power analyses (assuming a power of 0.80 and a two-tailed alpha of 0.05) indicated that the current sample size of 76 could have been insufficient to detect these significant effects as a sample size of 280 would have been needed to detect an association between confidence and Anxiety-Depression scores. It is also important to consider that findings from the general population may not always be generalisable to patient populations. A recent study comparing non-clinical highly compulsive individuals to OCD patients found that whereas highly compulsive individuals did indeed display local and global overconfidence, OCD patients exhibited underconfidence across all three levels of the metacognitive hierarchy (Hoven, Rouault, et al., 2022). This implies that confidence manifestations can significantly vary, even among populations sharing compulsive tendencies. Hence, drawing inferences from general population studies, such as Rouault, Seow, et al. (2018), about the way in which the transdiagnostic dimensions impact on a clinical group like problem gamblers should be done with caution. Although the symptom dimensions may be associated with confidence biases in such individuals, there could also be distinct aspects inherent to problem gamblers that modify the extent and manifestation of these biases.

Exploring the relationships between local trial confidence, global task confidence, and self-esteem, I found a strong association between local and global confidence, with no significant differences between the problem gamblers and control participants. This suggests that although problem gamblers are biased in their local confidence judgements for individual decisions, this information is then integrated into a global confidence judgement on a task level without further distortion. How-

ever, considering that global confidence was probed after each block of trials, and via a similar 6-point scale, it may not be surprising that this measure closely aligns with trial-level confidence. Interestingly, self-esteem appeared to be disconnected from both local and global confidence in the control participants, indicating a decoupling across the metacognitive hierarchy. In contrast to the control participants, there was a significant correlation between self-esteem and local confidence within the problem gamblers. Notably, the fact that the decoupling of self-esteem from local confidence was observed in the control participants is in contrast to my original hypothesis. This finding suggests that the dissociation is not likely driven by counteracting impacts of Compulsive Behaviour and Intrusive Thought on local confidence, and Anxiety-Depression on self-esteem. If this were the case, we would expect a stronger dissociation at higher symptom levels, i.e., in the problem gamblers. The observed independence of self-esteem from local and global confidence in control participants contrasts with recent research that revealed a positive association between individual confidence and self-esteem (Moses-Payne et al., 2019). Rouault et al. (2022) compared low and high self-esteem groups and discovered that, despite no significant performance disparity, the low self-esteem group consistently reported lower global confidence ratings. Corroborating this, Hoven et al. (2023) found that higher-order self-beliefs were positively correlated with confidence and overconfidence at both local and global levels, independent of objective performance. The apparent divergence of these findings from my results underscores the necessity for additional comprehensive, long-term studies, which would provide a more comprehensive understanding of the relationship between these variables and how they evolve over time.

It is noteworthy that although the transdiagnostic dimensions probed in this study have been validated repeatedly, the questionnaire items that constitute these dimensions do not comprehensively represent all forms of psychopathology. Other

transdiagnostic symptom structures that may capture a more extensive array of cognitive/metacognitive alterations have been suggested (Caspi et al., 2014; Keyes et al., 2013; Kotov et al., 2010). Furthermore, another fundamental question concerns the relationship between abnormalities in metacognitive processes and psychiatric disorders. These abnormalities might be intricately linked with, or even underpin, psychiatric symptoms, or they could arise as a consequence of the disorder. Alternatively, they might be inconsequential byproducts that have no significant influence on symptom presentation. In this context, Fox et al. (2023) found that the underconfidence bias related to Anxiety-Depression showed significant improvement along with reductions in Anxiety-Depression severity following cognitive-behavioural therapy or antidepressant medication. Although this finding does not clarify whether metacognitive abnormalities are a cause or consequence of the disorder, it offers valuable insight into their dynamic nature. Specifically, it suggests that metacognitive biases may not be enduring, static traits, but rather state-dependent variables susceptible to change as psychiatric symptoms evolve. However, more research is needed to fully understand the relationship between metacognitive abnormalities and psychiatric disorders.

Given that overconfidence in problem gamblers can lead to excessive risk-taking, increased financial loss, and a destructive cycle of continued gambling, a better understanding of the driving forces behind this overconfidence is needed to inform therapeutic interventions aimed at mitigating its adverse effects. The current study established that problem gamblers exhibit significantly higher levels of local and global decision confidence compared to a control group. Notably, this heightened decision confidence persists despite lower overall self-esteem and is not fully explained by the transdiagnostic symptom dimensions Compulsive Behaviour and Intrusive Thought and Anxiety-Depression. A future direction of this research might include a more comprehensive examination of cognitive flexibility and decision-making pro-

cesses in problem gamblers using gamified versions of other cognitive tasks. These tasks could provide additional insights into the cognitive profile of problem gamblers, contributing to a more nuanced understanding of the cognitive biases and distortions that may fuel overconfidence and persistent gambling behaviours in this group.

6

General Discussion

Chapter Abstract

This thesis has explored the role of confidence in adaptive decision-making. The first line of empirical research examined the flexibility with which confidence guides information seeking behaviour (Chapter 2). The second line of research investigated the impact of informational and normative social influences on confidence judgements (Chapters 3 and 4). The third line of research probed into confidence biases prevalent in problem gamblers (Chapter 5). Revisiting the core themes discussed in the introductory chapter — the role of confidence in guiding adaptive behaviours, its significance in social contexts, and its anomalies in psychiatric disorders — I briefly discuss each line of empirical work in its immediate context. I then consider how the findings come together in contributing to our understanding of confidence, and metacognition more generally, in a wider context. Here, I revisit some of the broader themes introduced in the opening chapter, considering our conceptualisation of metacognition, its potential neural representation, and its measurement in laboratory settings.

6.1 Metacognition guides adaptive behaviour

The opening chapter of this thesis highlighted the role of metacognition as a key facilitator in the adaptive and flexible regulation of behaviour. As reviewed, confidence, a metacognitive judgement, significantly influences various behaviours, from study choices (Hanczakowski et al., 2014; Nelson et al., 1994) to the setting of reminders (Gilbert, 2015; Risko and Gilbert, 2016) as well as guiding information seeking behaviours (Desender et al., 2018; Eliaz and Schotter, 2010; Pescetelli et al., 2021). Beyond these immediate applications, confidence also informs future strategic decisions by promoting caution after low-confidence decisions (Van den Berg et al., 2016), informing performance expectations (Boldt, Blundell, et al., 2019; Rouault, McWilliams, et al., 2018), steering future task selection (Carlebach and Yeung, 2020), and directing value-based decision-making (Boldt, Blundell, et al., 2019; Folke et al., 2016).

Much of the reviewed literature portrays a fixed relationship between confidence and behaviour: low confidence prompts one course of action, such as seeking information before committing to a choice, whereas high confidence triggers an alternative action, such as making an immediate decision without additional information. Although past research lends credibility to the central role that confidence plays in adaptively steering behaviour, this perspective falls short of embracing the complexity and diversity inherent in day-to-day circumstances. A key question that remains to be further explored is the flexibility with which confidence is used in guiding our actions. In fact, cognitive flexibility is a key characteristic of human cognition, enabling us to adapt in response to changing demands (Ionescu, 2012). Importantly, conscious representations like confidence are believed to play a critical role in this flexibility (Dehaene and Changeux, 2011).

Thus, in Chapter 2, I examined the flexibility with which confidence guides information seeking, postulating that the relationship between confidence and behaviour is not necessarily fixed, but may instead vary based on the specific context. The conducted experiments underscore the inherent flexibility in confidence-guided information seeking behaviour across varying contexts. In the first experiment, for comparable levels of confidence, participants sought more information in a task in which incorrect decisions led to losses compared to one in which the reward difference between correct and incorrect choices was minimal. The second experiment reiterated this flexibility, with participants adjusting their information seeking behaviour relative to the potential cost of revising an initial decision upon gathering more information in a given task block. The insights gleaned from both experiments underscore that participants modulate their information seeking behaviour responsive to the context.

Despite this evidence of flexibility, the experiments also revealed certain limitations on such adaptability. In the first experiment, I proposed that in a task in which incorrect decisions led to losses, participants would increase their information seeking on trials with high-value choice options, given the heightened stakes with high potential rewards and losses hinging on the correctness of choices. Moreover, in an task mirroring real-world value-based decisions — in which choices usually are not simply right or wrong, but rather fall on a continuum from better to worse — I predicted that participants would reduce information seeking when confidence was low. This inversion of the conventional confidence-information seeking relationship was premised on the assumption that low confidence signaled similar values among choice options, thereby making the quest for additional information to identify the slightly superior option not worthwhile on such trials. Contrary to these predictions, participants did not adjust their information seeking behaviour on a trial-by-trial basis. Instead, they defaulted to a heuristic strategy, increasing their

information seeking when confidence was lower, regardless of the potential reward or loss hinging on their choice. This pattern persisted in the second experiment, where again participants' information seeking behaviour remained unresponsive to the trial-by-trial value of choice options and consequently the reward at stake. Instead, participants consistently associated lower confidence with a higher need for information, suggesting a reliance on heuristic strategies over a nuanced evaluation of potential outcomes.

These findings suggest that although individuals exhibit a degree of flexibility in how confidence guides their behaviour, heuristic strategies prevail when it comes to more nuanced aspects of the choice environment. Participants appeared to weigh the costs and benefits of information seeking at the broader task or block level, yet resorted to a default rule at the trial level, whereby lower confidence generally signalled a greater need for additional information before finalising a choice. These results offer valuable insights that extend the literature on how confidence guides behaviour in an adaptive and context-dependent fashion (Carlebach and Yeung, 2023). Cognitive and behavioural flexibility enable the dynamic adjustment of thoughts and actions to appropriately meet the changing demands of the environment (Uddin, 2021). Despite having cognitive mechanisms in place that allow for adaptive flexibility, individuals often revert to default, simple rules to guide their decisions, balancing the benefits of accurate decisions with the cognitive costs associated with complex decision-making processes (Gigerenzer and Gaissmaier, 2011). The observation of this flexible, yet heuristically bound use of confidence to guide information seeking in my experiments, adds to our understanding of the way in which humans make decisions and adapt to their environment, and emphasises the integral role of metacognition in guiding our actions.

6.2 Metacognition in social settings

Decisions are rarely made in social isolation. Social interaction necessitates the flexibility of behaviour. Taking into account the viewpoints and motivations of others can have an impact on our own preferences (Garvert et al., 2015; Izuma and Adolphs, 2013), learning (Behrens et al., 2008; Boorman et al., 2013; Wittmann et al., 2016), decisions (Hill et al., 2017), and confidence (Bang et al., 2017; Kappes et al., 2020; Mahmoodi et al., 2019; Rousseau and Kashur, 2021; Will et al., 2017; Wittmann et al., 2016).

According to several theories, one of the key functions of metacognition lies in its role in facilitating the communication of our own cognitive processes to others (Frith, 2012; Frith and Frith, 2012; Shea et al., 2014). Confidence serves as a simple yet potent signal, conveying meaningful insights about the reliability of a choice or opinion, thus eliminating the need to elaborate on the intricate mechanisms of the decision-making process or the underlying supporting evidence. Confidence estimates, for example, play a significant part in seeking and giving advice (Price and Stone, 2004; Sniezek and Van Swol, 2001; Yates et al., 1996), and in judging the trustworthiness of eyewitness testimonies (Brewer and Hupfeld, 2004; Penrod and Cutler, 1995; Tenney et al., 2007). The expression of confidence often serves as a cue to the reliability of an opinion (Price and Stone, 2004), and sharing metacognitive evaluations aids in group decision-making and can enhance collective performance (Bahrami et al., 2013; Bahrami et al., 2012; Bahrami et al., 2010; Brennan and Enns, 2015; Fusaroli et al., 2012; Koriat, 2012; Pescetelli et al., 2016; Zarnoth and Sniezek, 1997).

In Chapters 3 and 4, I examined a phenomenon termed '*confidence matching*', a process by which individuals, when making decisions together, have a tendency to align

their confidence reports to one another. Building on classic findings regarding social conformity (Asch, 1951; Sherif, 1935), I hypothesised that this phenomenon may be driven by two main factors: informational influences that prompt individuals to truly adjust their confidence according to their partners' confidence, and normative influences that encourage individuals to alter the way in which they express their confidence outwardly, without changing their internal confidence estimate. Findings from Chapter 3 suggest that both factors can contribute to confidence matching. In an environment lacking feedback, participants aligned their confidence judgements with those of their partners and adjusted their information seeking behaviour — an indicator for internal confidence — accordingly, indicating the role of informational influences. Conversely, in a situation in which feedback was readily available but a collective decision-making rule was imposed, participants adapted their confidence judgements to match their partners' but did not alter their information seeking behaviour, suggesting that normative influences play a role in shaping the external manifestation of confidence matching. In common with findings from the previous line of research presented in Chapter 2, these results again underscore the adaptability and context-dependence of confidence judgements.

Whereas the experiments in Chapter 3 used information seeking behaviour as a proxy for internal confidence, the aim of Chapter 4 was to extend this by applying a neuroimaging-based approach. Here, I focused on single-trial electroencephalography (EEG) measures of neural activity related to internal confidence, as have been identified in previous research (Dou et al., 2023; Fernandez-Vargas et al., 2020; Sadras et al., 2022; Valeriani et al., 2022). In this EEG-based experiment, I hypothesised that both informational and normative social influences would result in overt confidence matching, albeit with distinct effects on the neural markers of confidence. Under informational influences, I anticipated that aligning confidence with a partner's reports would alter the EEG-based neural markers of confidence, whereas

under normative influences, I posited that such alignment would modify the relationship between EEG-based neural markers and overt confidence reports. However, as no reliable EEG marker of confidence was identified that could have been employed to test the hypotheses, the results from this study remain inconclusive.

In sum, the results from Chapters 3 and 4 build upon and extend the literature on the flexible use of confidence and its role in social situations. They highlight that individuals not only utilise confidence to navigate their personal decisions but also to coordinate their decisions with others, and that the social environment plays a crucial role in shaping metacognitive judgements. The findings suggest that the formation and expression of confidence judgements can be substantially shaped by both informational and normative influences within a group. The distinction of these two influences and their differential impacts on confidence matching adds granularity to our understanding of how metacognition operates in social interactions.

6.3 Metacognition in psychiatric disorders

In the introduction of this thesis, I reviewed how confidence judgements may be affected in various psychiatric conditions, with the nature of the observed confidence aberrations differing by disorder (Hoven et al., 2019). For instance, Obsessive-Compulsive Disorder (OCD) has been associated with diminished confidence levels, especially in OCD-related contexts like checking behaviour (Karadag et al., 2005; Lazarov et al., 2012; Macdonald et al., 1997; McNally and Kohlbeck, 1993; Moritz and Jaeger, 2018; Zhang et al., 2017). Moreover, individuals with depression typically exhibit decreased confidence levels (Fieker et al., 2016; Fu et al., 2005; Hancock, 1996; Stone et al., 2001), possibly mirroring the negative information processing bias seen in mood disorders (Gotlib and Krasnoperova, 1998). Conversely, schizophrenia is associated with heightened confidence, specifically in errors, leading to diminished

metacognitive sensitivity (Bhatt et al., 2010; Davies et al., 2018; Eifler et al., 2015; Gawęda et al., 2012; Kircher et al., 2007; Köther et al., 2012; Moritz and Woodward, 2002; Moritz et al., 2004; Moritz et al., 2003; Moritz et al., 2012; Peters et al., 2013; Warman, 2008), an observation that resonates with core symptoms of schizophrenia like delusions and hallucinations (Moritz et al., 2005). In addition, addiction is associated with heightened confidence and diminished metacognitive efficiency (Le Berre et al., 2010; Mintzer and Stitzer, 2002; Moeller et al., 2016; Sadeghi et al., 2017), possibly reflecting the lack of self-insight often seen in addiction (Goldstein et al., 2009). A comparable pattern of overconfidence is evident in gambling disorder (Brewer and Hupfeld, 2004; Goodie, 2005; Lakey et al., 2007).

The heterogeneity of symptom profiles within a single diagnosis (Newson et al., 2021) and the high rates of comorbidity between disorders (Lépine et al., 2005) pose challenges to our understanding of aberrant patterns in (meta)cognition. In response, computational psychiatry (Friston et al., 2014; Huys et al., 2016; Stephan and Mathys, 2014) strives to relate the cognitive processes underlying complex behaviours to transdiagnostic symptoms that bear significance for mental health. Within this transdiagnostic framework, Rouault, Seow, et al. (2018) identified a symptom dimension Anxiety-Depression to be associated with lower confidence, and a dimension comprising Compulsive Behaviour and Intrusive Thought to correlate with elevated confidence.

Building on this, in Chapter 5, I focused on exploring confidence biases in problem gamblers and their potential associations with these transdiagnostic symptom dimensions. My findings demonstrated that problem gamblers displayed a significant increase in local confidence for individual decisions and in global confidence for their overall task performance when compared to control participants. This was true despite problem gamblers having generally lower self-esteem and comparable objec-

tive task accuracy. This pattern of overconfidence persisted even after controlling for Anxiety-Depression and Compulsive Behaviour and Intrusive Thought symptom levels, both of which were elevated in problem gamblers compared to control participants.

My results contribute to and expand the existing body of research on metacognitive biases in psychopathology by offering empirical evidence for elevated confidence levels in problem gamblers. Although prior studies have pointed towards a trend for overconfidence in problem gamblers (Brevers et al., 2014; Goodie, 2005; Hoven, de Boer, et al., 2022; Lakey et al., 2007), these studies failed to adequately control for performance differences or to have a demographically matched control group. Additionally, my findings suggest that the mechanisms driving overconfidence in problem gambling are at least partially distinct from those that have been linked with transdiagnostic symptoms (Rouault, Seow, et al., 2018).

6.4 Refining the model of metacognition

At the outset of this thesis, metacognition was introduced as '*knowledge and cognition about cognitive phenomena*' (Flavell, 1979, p. 906), or thinking about thinking. This concept was further elaborated upon in a model of metacognition formulated by Nelson and Narens (1990). According to this model, higher-order cognitive processes at the meta-level monitor and control lower-level cognitive processes at the object-level (e.g., memory, perception, or decision-making) in a feedforward and feedback manner (see Figure 1.1).

The seminal work by Nelson and Narens (1990) delineates a foundational framework wherein metacognitive monitoring exists to enable the online control of our cognitive processes. Past research has shown that individuals are able to give confidence judgements that, although not perfect, predict objective accuracy in a reliable man-

ner (Baranski and Petrusic, 1994). This correlation with objective accuracy is the basis for the functional value of confidence. Throughout this thesis, supported by both literature review and empirical research, two key questions have been considered: (1) what kind of information do people consider and utilise when monitoring their cognitions and making confidence judgements based on their observations, and (2) how is metacognitive control exercised, and in what ways do these metacognitive evaluations inform behaviour across various domains. By doing so, the original framework has been expanded upon, revealing further complexities. The idea of the meta-level as merely mirroring monitoring experiences that are broadcast to the relevant cognitive control processes is gradually giving way to a more nuanced understanding.

A key emerging idea is the presence of overarching metacognitive beliefs alongside immediate metacognitive experiences. Thus, the meta-level may be better represented as a hierarchical structure. Overarching beliefs, though informed by individual experiences (e.g., your task performance estimate may be formed by aggregating multiple experiences of confidence on single trials within that task), are also instrumental in shaping the very experiences that form them (e.g., your interpretation of your experiences may depend on your higher-order beliefs). This idea is further articulated in the metacognitive hierarchy framework, which proposes several interconnected levels of metacognition, extending from local, immediate confidence judgements to more global, overarching beliefs about one's abilities (Seow et al., 2021). At the base of the hierarchy, lies '*local*' metacognition, which relates to our immediate self-evaluations and judgements about specific decisions. These immediate confidence judgements allow regulation of subsequent decisions, affecting behaviours such as information gathering (Desender et al., 2018), exploration (Donoso et al., 2014), and adaptation of speed-accuracy trade-offs (Van den Berg et al., 2016). However, this very immediate form of metacognition is limited in its scope and

time-frame. To account for longer-term thinking and broader self-perceptions, the concept of '*global*' metacognition is introduced. Global metacognitive evaluations or self-beliefs cover multiple decisions, allowing us to form broader estimates of our own performance and abilities (Boldt, Blundell, et al., 2019; Carlebach and Yeung, 2020; Rouault, McWilliams, et al., 2018). This premise of a metacognitive system capable of aggregating confidence over multiple decisions aligns with recent studies demonstrating a confidence '*leak*' across consecutive decisions (Aguilar-Lleyda et al., 2021; Kantner et al., 2019; Rahnev et al., 2015). These self-beliefs can span from specific tasks, through entire cognitive domains, to overall perceptions of capability. More broadly, one could perceive global confidence as an initial move towards bridging the conceptual divide between confidence in a single decision and higher-order constructs of self-confidence and self-efficacy (Bandura, 1977). These constructs bear profound implications for mental health as they affect individuals' self-perception and their appraisal of their own capabilities (Bandura, 1977; Sowislo and Orth, 2013; Wells, 2011).

Investigating the mechanisms of metacognition, researchers have designed experiments that link local confidence estimates to the formation of more global beliefs (Cavalan et al., 2023; Lee, de Gardelle, et al., 2021; Rouault et al., 2019; Rouault and Fleming, 2020). Yet, a noticeable discrepancy exists between the experimental investigations of global self-beliefs and those relevant to real-life decisions. Real-life self-beliefs are likely shaped by a more diverse set of influences and fluctuate over considerably longer timescales than those assessed in laboratory environments (Will et al., 2017). Reinforcing the idea that global beliefs extend beyond mere accumulations of local confidence experiences, studies have shown that, especially for knowledge-based decisions, global confidence across a series of decisions often falls below the mean of individual confidence ratings for each decision (Gigerenzer et al., 1991; Griffin and Tversky, 1992; Sniezek and Buckley, 1991; Treadwell and Nelson,

1996). This trend has been attributed by some researchers to the distinct cues used in forming local and global confidence judgements. Gigerenzer et al. (1991) proposed that individuals rely on different sets of probabilistic cues when crafting local and global confidence estimates. Griffin and Tversky (1992) similarly proposed that whereas local confidence primarily hinges on the strength and weight of evidence, global confidence also factors in considerations such as task difficulty or past experiences.

Conceptualising the meta-level in this way as more complex and allowing for beliefs and knowledge on top of metacognitive experiences, provides a mechanism by which many other influences aside from our moment-to-moment experiences inform our metacognitive monitoring judgements and our corresponding metacognitive control behaviour. For instance, in Chapter 5, I found problem gamblers to be more confident in individual decisions compared to controls. Yet their overall self-esteem was lower. The elevated decision confidence may stem from a diminished sensitivity to objective evidence (Hoven, de Boer, et al., 2022), cognitive distortions (Cowley et al., 2015; Goodie, 2005; Goodie and Fortune, 2013; Ledgerwood et al., 2020; Mallorquí-Bagué et al., 2019; Orgaz et al., 2013), and cognitive inflexibility (Perandrés-Gómez et al., 2021; Van Timmeren et al., 2018) affecting their confidence computation for single decisions. However, their overall self-esteem could be influenced by a broader array of factors that extend beyond individual decisions. Such influences may include negative feedback from their environment arising from strained interpersonal relationships, financial distress, or adverse career outcomes associated with their gambling behaviour.

Additionally, considering the meta-level in this way provides a means for communication of our metacognitive beliefs and thereby influencing and being influenced by others. Chapter 3 provided evidence that in the absence of feedback, partici-

pants' internal confidence judgements were influenced by those of social partners who made their choices alongside them. This suggests that observing a partner's confidence judgements might impact an internal prior on expected task performance, such that socially induced alterations in local trial confidence may have been a result of changes in a global self-performance estimate. This interpretation further supports the notion that overarching metacognitive beliefs, though formed by individual experiences, also shape the very experiences that form them. Moreover, the observation of both normative and informational social influences on confidence reports underscores that metacognitive monitoring judgements and corresponding metacognitive control behaviour are influenced not solely by individual metacognitive experiences, but also by additional factors such as social interactions with others.

This idea of a prior on confidence ties in with Bayesian models of cognitive processes. Work by Olawole-Scott and Yon (2023) suggests that our brains do not merely assess the reliability or precision of incoming sensory signals for perceptual inference and confidence formation. Instead, they also form expectations about the precision of these perceptions, which in turn guide metacognition and awareness. In the experiments conducted, participants were found to develop probabilistic expectations about the upcoming signal strength of visual motion stimuli. These expectations led participants to feel more confident and perceive stimuli as more vivid when they anticipated stronger sensory signals, irrespective of actual perceptual performance. Computational modelling corroborated these findings, illustrating this effect as a product of a predictive learning model, whereby the perceived precision of signals is a weighted combination of incoming evidence and top-down expectations. This implies that our cognitive processes not only interpret the reliability of incoming sensory data but also factor in our prior knowledge about the likely precision of different information sources. Similarly, Hu et al. (2021) introduced a Bayesian

Inference Model for meta-memory, suggesting that people evaluate their memory performance based on both their processing experience during the memory process and their prior beliefs about overall memory ability. Thus, these studies underscore the hierarchical nature of metacognition, suggesting that confidence judgements for individual decisions or memories are shaped by higher-level priors on confidence.

Likewise, the findings in Chapter 2 provide further illumination on this hierarchical approach to metacognition. Here, the participants' information seeking decisions — a form of metacognitive control — seemed to be influenced both by their immediate, local confidence pertaining to each trial and by their broader, global beliefs about the inherent value of information within the context of an entire task or block of trials. Specifically, the participants' confidence in each individual decision appeared to dictate their immediate information seeking behaviours, with lower confidence leading to increased information seeking in order to improve decision accuracy. On a broader scale, participants seemed to form certain beliefs or expectations about the value of information within the larger context of the entire task or block. These global beliefs may influence how individuals approach a given task and how they manage their information seeking behaviours. For instance, a participant who believes that a task is inherently difficult or that the rewards for accurate decisions are significant may prioritise information gathering more generally throughout the task. My findings however suggest that these two levels of metacognition — the local trial-level confidence and the global task-level beliefs — did not interact with each other in determining information seeking behaviours. Rather, they seemed to exert independent, additive effects. In other words, an individual's information seeking behaviour was affected by their confidence in their decision on each trial (the lower the confidence, the more information sought), and separately by their overall beliefs about the task (with higher perceived value of information leading to more information seeking). Although both immediate confidence and overarching beliefs

played significant roles in guiding information seeking choices in my experiments, they appeared to operate in parallel but non-interactive.

In conclusion, the studies presented in this thesis build upon the foundational framework of metacognition by Nelson and Narens (1990), and provide empirical evidence supporting the expansion of this model. We have seen that metacognition is not a simple mirror of monitoring experiences, but rather, a complex hierarchical structure that allows for the influence of factors beyond individual decision-making experiences to shape our metacognitive monitoring and control, shape the way that these metacognitive evaluations guide adaptive decision-making, and provide a mechanism through which we can communicate our metacognitive experiences with others. The Nelson and Narens (1990) model, as originally conceptualised, represents metacognition in a rather simplified manner: a circular entity that oversees and orchestrates our cognitive processes. Though revolutionary for its time, it inadvertently poses the homunculus problem, hinting at an internal entity in the mind that oversees cognitive reflection without specifying its operational mechanisms. One of the primary endeavors of metacognition research, exemplified in this thesis, is to systematically untangle this homunculus problem, by breaking down and specifying the inherent mechanisms of metacognition. By formalising these parts, we are in essence deconstructing the metaphorical circle, specifying its mechanisms, and, in the process, eliminating the ambiguous homuncular nature. The implications of such an expansion are manifold. For example, by discerning between the multiple layers of metacognition,we open the door to more intricate investigations into the nature of metacognitive processes: For instance, are local and global metacognition underpinned by separate cognitive or neural systems, and how do these identified levels interrelate and interface with each other? By breaking down metacognition into distinct components, we pave the way for a clearer, more mechanistic understanding of how we think about our thinking, laying the foundation for a plethora

of investigative opportunities in the realm of metacognitive research.

6.5 Neural representation of confidence

In the opening chapter, it became apparent that there are inconsistencies and ongoing debate in the current literature surrounding the neural underpinnings of metacognition (e.g., Boldt and Yeung, 2015; Gherman and Philiastides, 2018; Hebart et al., 2016; Kepecs et al., 2008; Komura et al., 2013; Lak et al., 2017; Middlebrooks and Sommer, 2012; Miyamoto et al., 2017; Vaccaro and Fleming, 2018). This prompted the suggestion of three main theories positing how confidence might be instantiated in the brain. The Canonical Model of Metacognition conceives confidence as an abstract, universal representation. This model purports that confidence serves as a bridge between varying types of inputs, such as confidence within different cognitive domains or sensory modalities, and varying types of outputs, such as confidence applied for decision-making or interpersonal communication. Alternatively, the Domain-Specific Metacognition Model proposes distinct confidence representations corresponding to different cognitive domains, such as memory or perception. These individual representations, despite being distinct, might contain shared attributes that lead to similar subjective experiences. Lastly, the Emergent Metacognition Model challenges traditional perspectives, rejecting the notion of confidence as an abstract construct synthesizing multiple cues into a singular representation. Instead, it proposes that individual cues, such as fluency or familiarity, directly lead to strategic adaptations, like information seeking or reminder setting.

Although my empirical research did not explicitly test contrasting predictions emerging from these models, the insights generated could still hold relevance for this discourse. Experiments in Chapter 2 showed that participants could utilise their confidence in a flexible manner, modulating how confidence guides information seeking

across distinct task environments. Furthermore, they used their confidence to generalise within a task environment, heuristically resorting to confidence as an indicator of the need for more information prior to making a decision. As such, two main attributes of confidence emerged from the research: flexibility between tasks and generalisation within a task. A more abstract general representation of confidence would more easily allow for this kind of flexibility and generalisation.

Experiments in Chapter 3 showed that under normative social influences, participants exhibited flexibility in using their confidence estimates to guide information seeking choices independent of how they mapped their confidence onto overt reports depending on the social context. This implies that participants were able to separate how their confidence related to different actions, i.e., confidence used for information seeking versus confidence used to generate a socially suitable report. Moreover, the ability and need for communication of confidence presents a strong case for an explicit general representation over the direct effect of various cues on behavioural adaptations.

In the experiment outlined in Chapter 5, problem gamblers exhibited overconfidence in a simple perceptual decision task, unrelated to gambling and devoid of any monetary reward. The finding that this overconfidence bias persists across contexts lends support to at least some kind of domain-generality in the way in which confidence is computed.

In conclusion, my research aligns with models that highlight the significance of explicit conscious representations in driving flexible, goal-directed behaviour (Dehaene and Changeux, 2011; Nelson and Narens, 1990; Shea and Frith, 2019). Notably, having an explicit representation of confidence, making it accessible across various cognitive systems, offers distinct advantages. It facilitates planning and reasoning within as well as across individuals, incorporating the reliability of the information

at hand (Shea et al., 2014; Shea and Frith, 2019). Moreover, it enables generalisation across contexts, allowing the same explicit representation to trigger an adaptive response in varying situations without the need to learn each association anew.

6.6 How we ask matters

The initial chapter of this thesis introduced various strategies for eliciting metacognitive judgements from participants in a controlled lab environment. Throughout this thesis, it has become progressively clear that the methodology of inquiry significantly influences the results and, consequently, their interpretation. Nuanced considerations surrounding the process of eliciting metacognitive judgements can influence the participants' responses. For example, as highlighted in Chapter 2, participants indicated higher confidence levels when dealing with higher-value choice options. This raises an important question: Does this change reflect a genuine increase in confidence, or are participants misinterpreting our confidence scale, viewing it more as a measure of expected outcome quality rather than an assessment of the accuracy of their perceptual decisions? Adding to the complexity, results from Chapters 3 and 4 indicate that social contexts significantly influence how participants report their confidence levels. Specifically, I found that people implicitly or explicitly follow different social norms for how confidence should be communicated. A related concern is the issue of reactivity—changes in behaviour that arise when individuals are aware that they are being observed. Reactivity can take on various forms, from alterations in motivation and expectations to shifts in emotions and goals, and it has been shown to affect measures of metacognition as well (Double and Birney, 2019; Double et al., 2018). Given the influence of our questioning techniques on eliciting metacognitive judgements, it may be advantageous to use more implicit measures. For example, utilising strategies such as information seeking behaviour as a proxy for confidence or examining neural correlates of confidence might provide

useful alternatives or complements to self-report scales.

6.7 Conclusion

The capacity to reflect on our own thoughts and behaviour is a key part of human cognition. The empirical work presented in this thesis aimed to investigate the role of confidence in adaptive behaviour. As the results reveal, metacognition is not a monolithic construct but rather a complex, context-sensitive mechanism that directs our actions and takes into consideration a range of influences that extend beyond immediate experiences. The results highlight the important role that the study of metacognition, and specifically confidence, plays in multiple facets of human life, from everyday decision-making to clinical diagnosis and treatment.

Bibliography

Aguilar-Lleyda, D., Konishi, M., Sackur, J., & de Gardelle, V. (2021). Confidence can be automatically integrated across two visual decisions. *Journal of Experimental Psychology: Human Perception and Performance*, *47*(2), 161.

Ais, J., Zylberberg, A., Barttfeld, P., & Sigman, M. (2016). Individual consistency in the accuracy and distribution of confidence judgments. *Cognition*, *146*, 377–386.

Allen, M., Frank, D., Schwarzkopf, D. S., Fardo, F., Winston, J. S., Hauser, T. U., & Rees, G. (2016). Unexpected arousal modulates the influence of sensory noise on confidence. *Elife*, *5*, e18103.

American Psychiatric Association. (2022). *Diagnostic and statistical manual of mental disorders* (5, text rev.). https://doi.org/10.1176/appi.books.9780890425787

American Psychiatric Association, D., Association, A. P., et al. (2013). *Diagnostic and statistical manual of mental disorders: Dsm-5* (Vol. 5). American psychiatric association Washington, DC.

Arbuckle, T. Y., & Cuddy, L. L. (1969). Discrimination of item strength at time of presentation. *Journal of experimental psychology*, *81*(1), 126.

Ariel, R., Lembeck, N. A., Moffat, S., & Hertzog, C. (2018). Are there sex differences in confidence and metacognitive monitoring accuracy for everyday, academic, and psychometrically measured spatial ability? *Intelligence*, *70*, 42–51.

Arnold, D. H., Johnston, A., Adie, J., & Yarrow, K. (2023). On why we lack confidence in some signal-detection-based analyses of confidence. *Consciousness and Cognition*, *113*, 103532.

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. *Groups, leadership, and men*, 177–190.

Bahrami, B., Didino, D., Frith, C., Butterworth, B., & Rees, G. (2013). Collective enumeration. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(2), 338.

Bahrami, B., Olsen, K., Bang, D., Roepstorff, A., Rees, G., & Frith, C. (2012). Together, slowly but surely: The role of social interaction and feedback on the build-up of benefit in collective decision-making. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(1), 3.

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science*, *329*(5995), 1081–1085.

Baird, B., Cieslak, M., Smallwood, J., Grafton, S. T., & Schooler, J. W. (2015). Regional white matter variation associated with domain-specific metacognitive accuracy. *Journal of cognitive neuroscience*, *27*(3), 440–452.

Baird, B., Smallwood, J., Gorgolewski, K. J., & Margulies, D. S. (2013). Medial and lateral networks in anterior prefrontal cortex support metacognitive ability for memory and perception. *Journal of Neuroscience*, *33*(42), 16657–16665.

Baker, L. (1984). Metacognitive skills and reading. *Handbook of reading research*.

Balsdon, T., Mamassian, P., & Wyart, V. (2021). Separable neural signatures of confidence during perceptual decisions. *Elife*, *10*, e68491.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological review*, *84*(2), 191.

Bandura, A. (1978). Self-efficacy: Toward a unifying theory of behavioral change. *Advances in behaviour research and therapy*, *1*(4), 139–161.

Bang, D., Aitchison, L., Moran, R., Herce Castanon, S., Rafiee, B., Mahmoodi, A., Lau, J. Y., Latham, P. E., Bahrami, B., & Summerfield, C. (2017). Confidence matching in group decision-making. *Nature Human Behaviour*, *1*(6), 0117.

Bang, D., Ershadmanesh, S., Nili, H., & Fleming, S. M. (2020). Private–public mappings in human prefrontal cortex. *Elife*, *9*, e56477.

Bang, D., & Fleming, S. M. (2018). Distinct encoding of decision confidence in human medial prefrontal cortex. *Proceedings of the National Academy of Sciences*, *115*(23), 6082–6087.

Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception & psychophysics*, *55*(4), 412–428.

Barron, H. C., Garvert, M. M., & Behrens, T. E. (2015). Reassessing vmpfc: Full of confidence? *Nature neuroscience*, *18*(8), 1064–1066.

Bartz, J. A., & Hollander, E. (2006). Is obsessive–compulsive disorder an anxiety disorder? *Progress in neuro-psychopharmacology and biological psychiatry*, *30*(3), 338–352.

Baumgarten, T. J., Schnitzler, A., & Lange, J. (2016). Prestimulus alpha power influences tactile temporal perceptual discrimination and confidence in decisions. *Cerebral Cortex*, *26*(3), 891–903.

Becker, M. P., Nitsch, A. M., Miltner, W. H., & Straube, T. (2014). A single-trial estimation of the feedback-related negativity and its relation to bold responses in a time-estimation task. *Journal of Neuroscience*, *34*(8), 3005–3012.

Behrens, T. E., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. (2008). Associative learning of social value. *Nature*, *456*(7219), 245–249.

Benwell, C. S., Mohr, G., Wallberg, J., Kouadio, A., & Ince, R. A. (2022). Psychiatrically relevant signatures of domain-general decision-making and metacognition in the general population. *npj Mental Health Research*, *1*(1), 10.

Bhatt, R., Laws, K. R., & McKenna, P. J. (2010). False memory in schizophrenia patients with and without delusions. *Psychiatry Research*, *178*(2), 260–265.

BioSemi, B. (2011). Biosemi activetwo.[eeg system]. *Amsterdam: BioSemi*.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual review of psychology*, *64*, 417–444.

Boldt, A., Blundell, C., & De Martino, B. (2019). Confidence modulates exploration and exploitation in value-based learning. *Neuroscience of consciousness*, *2019*(1), niz004.

Boldt, A., De Gardelle, V., & Yeung, N. (2017). The impact of evidence reliability on sensitivity and bias in decision confidence. *Journal of experimental psychology: human perception and performance*, *43*(8), 1520.

Boldt, A., & Gilbert, S. (2020). Distinct and overlapping neural correlates of metacognitive monitoring and metacognitive control. *Methods*, *2239*, 15.

Boldt, A., & Gilbert, S. J. (2022). Partially overlapping neural correlates of metacognitive monitoring and metacognitive control. *Journal of Neuroscience*, *42*(17), 3622–3635.

Boldt, A., Schiffer, A.-M., Waszak, F., & Yeung, N. (2019). Confidence predictions affect performance confidence and neural preparation in perceptual decision making. *Scientific reports*, *9*(1), 4031.

Boldt, A., & Yeung, N. (2015). Shared neural markers of decision confidence and error detection. *Journal of Neuroscience*, *35*(8), 3478–3484.

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational behavior and human decision processes*, *101*(2), 127–151.

Boorman, E. D., O'Doherty, J. P., Adolphs, R., & Rangel, A. (2013). The behavioral and neural mechanisms underlying the tracking of expertise. *Neuron*, *80*(6), 1558–1571.

Boschen, M. J., & Vuksanovic, D. (2007). Deteriorating memory confidence, responsibility perceptions and repeated checking: Comparisons in ocd and control samples. *Behaviour research and therapy*, *45*(9), 2098–2109.

Brennan, A. A., & Enns, J. T. (2015). When two heads are better than one: Interactive versus independent benefits of collaborative cognition. *Psychonomic Bulletin & Review*, *22*, 1076–1082.

Brevers, D., Cleeremans, A., Bechara, A., Greisen, M., Kornreich, C., Verbanck, P., & Noël, X. (2014). Impaired metacognitive capacities in individuals with problem gambling. *Journal of gambling studies*, *30*, 141–152.

Brewer, N., & Hupfeld, R. M. (2004). Effects of testimonial inconsistencies and witness group identity on mock-juror judgments 1. *Journal of Applied Social Psychology*, *34*(3), 493–513.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, *78*(1), 1–3.

Broihanne, M.-H., Merli, M., & Roger, P. (2014). Overconfidence, risk perception and the risk-taking behavior of finance professionals. *Finance Research Letters*, *11*(2), 64–73.

Bronson, M. (2000). Self-regulation in early childhood: Nature and nurture. new york: A division of guilford publications.

Brus, J., Aebersold, H., Grueschow, M., & Polania, R. (2021). Sources of confidence in value-based choice. *Nature communications*, *12*(1), 7337.

Bustamante, L., Lieder, F., Musslick, S., Shenhav, A., & Cohen, J. (2021). Learning to overexert cognitive control in a stroop task. *Cognitive, Affective, & Behavioral Neuroscience*, *21*, 453–471.

Campbell, W. K., Goodie, A. S., & Foster, J. D. (2004). Narcissism, confidence, and risk attitude. *Journal of behavioral decision making*, *17*(4), 297–311.

Carlebach, N., & Yeung, N. (2020). Subjective confidence acts as an internal cost-benefit factor when choosing between tasks. *Journal of Experimental Psychology: Human Perception and Performance*, *46*(7), 729.

Carlebach, N., & Yeung, N. (2023). Flexible use of confidence to guide advice requests. *Cognition*, *230*, 105264.

Carlson, T. A., Grootswagers, T., & Robinson, A. K. (2019). An introduction to time-resolved decoding analysis for m/eeg. *arXiv: Neurons and Cognition*.

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., Meier, M. H., Ramrakha, S., Shalev, I., Poulton, R., et al. (2014). The p factor: One general psychopathology factor in the structure of psychiatric disorders? *Clinical psychological science*, *2*(2), 119–137.

Cavalan, Q., Vergnaud, J.-C., & De Gardelle, V. (2023). From local to global estimations of confidence in perceptual decisions. *Documents de travail du Centre d'Économie de la Sorbonne*.

Chen, J., Wu, Y., Tong, G., Guan, X., & Zhou, X. (2012). Erp correlates of social conformity in a line judgment task. *BMC neuroscience*, *13*, 1–10.

Costermans, J., Lories, G., & Ansay, C. (1992). Confidence level and feeling of knowing in question answering: The weight of inferential processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(1), 142.

Cowley, E., Briley, D. A., & Farrell, C. (2015). How do gamblers maintain an illusion of control? *Journal of Business Research*, *68*(10), 2181–2188.

Cuttler, C., Sirois-Delisle, V., Alcolado, G. M., Radomsky, A. S., & Taylor, S. (2013). Diminished confidence in prospective memory causes doubts and urges to check. *Journal of behavior therapy and experimental psychiatry*, *44*(3), 329–334.

Daffner, K. R., Chong, H., Sun, X., Tarbi, E. C., Riis, J. L., McGinnis, S. M., & Holcomb, P. J. (2011). Mechanisms underlying age-and performance-related differences in working memory. *Journal of cognitive neuroscience*, *23*(6), 1298–1314.

Dannon, P. N., Lowengrub, K., Aizer, A., & Kotler, M. (2006). Pathological gambling: Comorbid psychiatric diagnoses in patients and their families. *Israel Journal of Psychiatry and Related Sciences*, *43*(2), 88.

David, A. S., Bedford, N., Wiffen, B., & Gilleen, J. (2014). Failures of metacognition and lack of insight in neuropsychiatric disorders. *The cognitive neuroscience of metacognition*, 345–365.

Davies, G., Rae, C. L., Garfinkel, S. N., Seth, A. K., Medford, N., Critchley, H. D., & Greenwood, K. (2018). Impairment of perceptual metacognitive accuracy and reduced prefrontal grey matter volume in first-episode psychosis. *Cognitive Neuropsychiatry*, *23*(3), 165–179.

De Gardelle, V., Le Corre, F., & Mamassian, P. (2016). Confidence as a common currency between vision and audition. *Plos one*, *11*(1), e0147901.

De Gardelle, V., & Mamassian, P. (2014). Does confidence use a common currency across two visual tasks? *Psychological science*, *25*(6), 1286–1288.

De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nature neuroscience*, *16*(1), 105–110.

Dehaene, S., & Changeux, J.-P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, *70*(2), 200–227.

Desender, K., Boldt, A., Verguts, T., & Donner, T. H. (2019). Confidence predicts speed-accuracy tradeoff for subsequent decisions. *Elife*, *8*, e43499.

Desender, K., Boldt, A., & Yeung, N. (2018). Subjective confidence predicts information seeking in decision making. *Psychological science*, *29*(5), 761–778.

Desender, K., Murphy, P., Boldt, A., Verguts, T., & Yeung, N. (2019). A postdecisional neural marker of confidence predicts information-seeking in decision-making. *Journal of Neuroscience*, *39*(17), 3309–3319.

Desender, K., Vermeylen, L., & Verguts, T. (2022). Dynamic influences on static measures of metacognition. *Nature communications*, *13*(1), 4208.

Donoso, M., Collins, A. G., & Koechlin, E. (2014). Foundations of human reasoning in the prefrontal cortex. *Science*, *344*(6191), 1481–1486.

Dou, W., Martinez Arango, L. J., Castaneda, O. G., Arellano, L., Mcintyre, E., Yballa, C., & Samaha, J. (2023). Neural signatures of evidence accumulation encode subjective perceptual confidence. *bioRxiv*, 2023–04.

Double, K. S., & Birney, D. P. (2019). Reactivity to measures of metacognition. *Frontiers in Psychology*, *10*, 2755.

Double, K. S., Birney, D. P., & Walker, S. A. (2018). A meta-analysis and systematic review of reactivity to judgements of learning. *Memory*, *26*(6), 741–750.

Dowling, N. A., Cowlishaw, S., Jackson, A. C., Merkouris, S. S., Francis, K. L., & Christensen, D. R. (2015). Prevalence of psychiatric co-morbidity in treatment-seeking problem gamblers: A systematic review and meta-analysis. *Australian & New Zealand Journal of Psychiatry*, *49*(6), 519–539.

Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N., & Pouget, A. (2012). The cost of accumulating evidence in perceptual decision making. *Journal of Neuroscience*, *32*(11), 3612–3628.

Du, X., Ren, Y., Wu, S., & Wu, Y. (2019). The impact of advice distance on advice taking: Evidence from an erp study. *Neuropsychologia*, *129*, 56–64.

Dune, S., Flynn, C., Sholdis, J., et al. (2017). 2016 northern ireland gambling prevalence survey.

Eifler, S., Rausch, F., Schirmbeck, F., Veckenstedt, R., Mier, D., Esslinger, C., Englisch, S., Meyer-Lindenberg, A., Kirsch, P., & Zink, M. (2015). Metamemory in schizophrenia: Retrospective confidence ratings interact with neurocognitive deficits. *Psychiatry Research*, *225*(3), 596–603.

Eisenacher, S., Rausch, F., Ainser, F., Mier, D., Veckenstedt, R., Schirmbeck, F., Lewien, A., Englisch, S., Andreou, C., Moritz, S., et al. (2015). Investigation of metamemory functioning in the at-risk mental state for psychosis. *Psychological medicine*, *45*(15), 3329–3340.

El Zein, M., Bahrami, B., & Hertwig, R. (2019). Shared responsibility in collective decisions. *Nature human behaviour*, *3*(6), 554–559.

Eliaz, K., & Schotter, A. (2010). Paying for confidence: An experimental study of the demand for non-instrumental information. *Games and Economic Behavior*, *70*(2), 304–324.

Faivre, N., Filevich, E., Solovey, G., Kühn, S., & Blanke, O. (2018). Behavioral, modeling, and electrophysiological evidence for supramodality in human metacognition. *Journal of Neuroscience*, *38*(2), 263–277.

Falkenstein, M., Hohnsbein, J., Hoormann, J., & Blanke, L. (1991). Effects of crossmodal divided attention on late erp components. ii. error processing in choice reaction tasks. *Electroencephalography and clinical neurophysiology*, *78*(6), 447–455.

Fernandez-Vargas, J., Valeriani, D., Cinel, C., Sadras, N., Ahmadipour, P., Shanechi, M. M., Citi, L., & Poli, R. (2020). Confidence prediction from eeg recordings in a multisensory environment. *Proceedings of the 2020 10th International Conference on Biomedical Engineering and Technology*, 269–275.

Ferris, J. A., & Wynne, H. J. (2001). *The canadian problem gambling index*. Canadian Centre on substance abuse Ottawa, ON.

Feuerriegel, D., Murphy, M., Konski, A., Mepani, V., Sun, J., Hester, R., & Bode, S. (2022). Electrophysiological correlates of confidence differ across correct and erroneous perceptual decisions. *NeuroImage*, *259*, 119447.

Fieker, M., Moritz, S., Köther, U., & Jelinek, L. (2016). Emotion recognition in depression: An investigation of performance and response confidence in adult female patients with depression. *Psychiatry research*, *242*, 226–232.

Fitzgerald, L. M., Arvaneh, M., Carton, S., O'Keeffe, F., Delargy, M., & Dockree, P. M. (2022). Impaired metacognition and reduced neural signals of decision confidence in adults with traumatic brain injury. *Neuropsychology*.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, *34*(10), 906.

Fleck, M. S., Daselaar, S. M., Dobbins, I. G., & Cabeza, R. (2006). Role of prefrontal and anterior cingulate regions in decision-making processes shared by memory and nonmemory tasks. *Cerebral Cortex*, *16*(11), 1623–1630.

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general bayesian framework for metacognitive computation. *Psychological review*, *124*(1), 91.

Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1338–1349.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in human neuroscience*, *8*, 443.

Fleming, S. M., Ryu, J., Golfinos, J. G., & Blackmon, K. E. (2014). Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain*, *137*(10), 2811–2822.

Fleming, S. M., Van Der Putten, E. J., & Daw, N. D. (2018). Neural mediators of changes of mind about perceptual decisions. *Nature neuroscience*, *21*(4), 617–624.

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science*, *329*(5998), 1541–1543.

Foa, E. B., Huppert, J. D., Leiberg, S., Langner, R., Kichic, R., Hajcak, G., & Salkovskis, P. M. (2002). The obsessive-compulsive inventory: Development and validation of a short version. *Psychological assessment*, *14*(4), 485.

Folke, T., Jacobsen, C., Fleming, S. M., & De Martino, B. (2016). Explicit representation of confidence informs future value-based decisions. *Nature Human Behaviour*, *1*(1), 0002.

Fowle, H. J., & Boschen, M. J. (2011). The impact of compulsive cleaning on confidence in memory and cleanliness. *Journal of anxiety disorders*, *25*(2), 237–243.

Fox, C. A., Lee, C. T., Hanlon, A., Seow, T., Lynch, K., Harty, S., Richards, D., Palacios, J., O'Keane, V., Stephan, K. E., et al. (2023). Metacognition in anxious-depression is state-dependent: An observational treatment study.

Fox, C. A., McDonogh, A., Tighe, D., Hanlon, A. K., & Gillan, C. M. (2022). Validation of a gamified smartphone assessment of metacognition: Associations with transdiagnostic psychiatric dimensions [Poster presented at the meeting of the British Association for Psychopharmacology Summer Conference, London, UK].

Friedman, N. P., & Robbins, T. W. (2021). The role of prefrontal cortex in cognitive control and executive function. *Neuropsychopharmacology*, 1–18.

Friston, K. J., Stephan, K. E., Montague, R., & Dolan, R. J. (2014). Computational psychiatry: The brain as a phantastic organ. *The Lancet Psychiatry*, *1*(2), 148–158.

Frith, C. D. (2012). The role of metacognition in human social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1599), 2213–2223.

Frith, C. D., & Frith, U. (2012). Mechanisms of social cognition. *Annual review of psychology*, *63*, 287–313.

Frömer, R., Lin, H., Dean Wolf, C., Inzlicht, M., & Shenhav, A. (2021). Expectations of reward and efficacy guide cognitive control allocation. *Nature Communications*, *12*(1), 1030.

Fu, T., Koutstaal, W., Fu, C. H., Poon, L., & Cleare, A. J. (2005). Depression, confidence, and decision: Evidence against depressive realism. *Journal of Psychopathology and Behavioral Assessment*, *27*, 243–252.

Fu, T. S.-T., Koutstaal, W., Poon, L., & Cleare, A. J. (2012). Confidence judgment in depression and dysphoria: The depressive realism vs. negativity hypotheses. *Journal of behavior therapy and experimental psychiatry*, *43*(2), 699–704.

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., & Tylén, K. (2012). Coming to terms: Quantifying the benefits of linguistic coordination. *Psychological science*, *23*(8), 931–939.

Gaertig, C., & Simmons, J. P. (2023). Are people more or less likely to follow advice that is accompanied by a confidence interval? *Journal of Experimental Psychology: General*.

Gaeta, H., Friedman, D., & Hunt, G. (2003). Stimulus characteristics and task category dissociate the anterior and posterior aspects of the novelty p3. *Psychophysiology*, *40*(2), 198–208.

Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychonomic bulletin & review*, *10*(4), 843–876.

Garner, D. M., Olmsted, M. P., Bohr, Y., & Garfinkel, P. E. (1982). The eating attitudes test: Psychometric features and clinical correlates. *Psychological medicine*, *12*(4), 871–878.

Garvert, M. M., Moutoussis, M., Kurth-Nelson, Z., Behrens, T. E., & Dolan, R. J. (2015). Learning-induced plasticity in medial prefrontal cortex predicts preference malleability. *Neuron*, *85*(2), 418–428.

Gawęda, Ł., Moritz, S., & Kokoszka, A. (2012). Impaired discrimination between imagined and performed actions in schizophrenia. *Psychiatry Research*, *195*(1-2), 1–8.

Gehring, W. J., & Willoughby, A. R. (2002). The medial frontal cortex and the rapid processing of monetary gains and losses. *Science*, *295*(5563), 2279–2282.

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

Gherman, S., & Philiastides, M. G. (2018). Human vmpfc encodes early signatures of confidence in perceptual decisions. *Elife*, *7*, e38293.

Giardini, F., Coricelli, G., Joffily, M., & Sirigu, A. (2008). Overconfidence in predictions as an effect of desirability bias. *Advances in decision making under risk and uncertainty*, 163–180.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual review of psychology*, *62*, 451–482.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A brunswikian theory of confidence. *Psychological review*, *98*(4), 506.

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. Oxford University Press, USA.

Gilbert, S. J. (2015). Strategic use of reminders: Influence of both domain-general and task-specific metacognitive confidence, independent of objective memory ability. *Consciousness and Cognition*, *33*, 245–260.

Gillan, C. M., Kalanthroff, E., Evans, M., Weingarden, H. M., Jacoby, R. J., Gershkovich, M., Snorrason, I., Campeas, R., Cervoni, C., Crimarco, N. C., et al. (2020). Comparison of the association between goal-directed planning and self-reported compulsivity vs obsessive-compulsive disorder diagnosis. *JAMA psychiatry*, *77*(1), 77–85.

Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A., & Daw, N. D. (2016). Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *elife*, *5*, e11305.

Gluth, S., Rieskamp, J., & Büchel, C. (2012). Deciding when to decide: Time-variant sequential sampling models explain the emergence of value-based decisions in the human brain. *Journal of Neuroscience*, *32*(31), 10686–10698.

Gold, J. I., & Shadlen, M. N. (2001). Neural computations that underlie decisions about sensory stimuli. *Trends in cognitive sciences*, *5*(1), 10–16.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annu. Rev. Neurosci.*, *30*, 535–574.

Goldstein, R. Z., Bechara, A., Garavan, H., Childress, A. R., Paulus, M. P., Volkow, N. D., et al. (2009). The neurocircuitry of impaired insight in drug addiction. *Trends in cognitive sciences*, *13*(9), 372–380.

Goodie, A. S. (2005). The role of perceived control and overconfidence in pathological gambling. *Journal of gambling studies*, *21*, 481–502.

Goodie, A. S., & Fortune, E. E. (2013). Measuring cognitive distortions in pathological gambling: Review and meta-analyses. *Psychology of Addictive Behaviors*, *27*(3), 730.

Gotlib, I. H., & Krasnoperova, E. (1998). Biased information processing as a vulnerability factor for depression. *Behavior Therapy*, *29*(4), 603–617.

Goudriaan, A. E., Oosterlaan, J., de Beurs, E., & van den Brink, W. (2005). Decision making in pathological gambling: A comparison between pathological gamblers, alcohol dependents, persons with tourette syndrome, and normal controls. *Cognitive brain research*, *23*(1), 137–151.

Goudriaan, A. E., van den Brink, W., & van Holst, R. J. (2019). Gambling disorder and substance-related disorders: Similarities and differences. *Gambling disorder*, 247–269.

Goupil, L., & Kouider, S. (2019). Developing a reflective mind: From core metacognition to explicit self-reflection. *Current Directions in Psychological Science*, *28*(4), 403–408.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., et al. (2013). Meg and eeg data analysis with mne-python. *Frontiers in neuroscience*, 267.

Green, J., Goldstein, F. C., Sirockman, B. E., & Green, R. C. (1993). Variable awareness of deficits in alzheimer's disease. *Neuropsychiatry, Neuropsychology, & Behavioral Neurology*.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive psychology*, *24*(3), 411–435.

Grogan, J. P., Rys, W., Kelly, S. P., & O'Connell, R. G. (2023). Confidence is predicted by pre-and post-choice decision signal dynamics. *bioRxiv*, 2023–01.

Guggenmos, M. (2021). Measuring metacognitive performance: Type 1 performance dependence and test-retest reliability. *Neuroscience of consciousness*, *2021*(1), niab040.

Guggenmos, M., Wilbertz, G., Hebart, M. N., & Sterzer, P. (2016). Mesolimbic confidence signals guide perceptual learning in the absence of external feedback. *Elife*, *5*, e13388.

Hancock, J. (1996). "depressive realism" assessed via confidence in decision-making. *Cognitive neuropsychiatry*, *1*(3), 213–220.

Hanczakowski, M., Zawadzka, K., & Cockcroft-McKay, C. (2014). Feeling of knowing and restudy choices. *Psychonomic bulletin & review*, *21*(6), 1617–1622.

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of educational psychology*, *56*(4), 208.

Hauser, T. U., Allen, M., Rees, G., & Dolan, R. J. (2017). Metacognitive impairments extend perceptual decision making weaknesses in compulsivity. *Scientific reports*, *7*(1), 6614.

Hebart, M. N., Schriever, Y., Donner, T. H., & Haynes, J.-D. (2016). The relationship between perceptual decision variables and confidence in the human brain. *Cerebral cortex*, *26*(1), 118–130.

Herding, J., Ludwig, S., von Lautz, A., Spitzer, B., & Blankenburg, F. (2019). Centro-parietal eeg potentials index subjective evidence and confidence during perceptual decision making. *NeuroImage*, *201*, 116011.

Hill, C. A., Suzuki, S., Polania, R., Moisa, M., O'doherty, J. P., & Ruff, C. C. (2017). A causal account of the brain network computations underlying strategic social behavior. *Nature neuroscience*, *20*(8), 1142–1149.

Holroyd, C. B., & Coles, M. G. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological review*, *109*(4), 679.

Holroyd, C. B., Pakzad-Vaezi, K. L., & Krigolson, O. E. (2008). The feedback correct-related positivity: Sensitivity of the event-related brain potential to unexpected positive feedback. *Psychophysiology*, *45*(5), 688–697.

Hoven, M., de Boer, N. S., Goudriaan, A. E., Denys, D., Lebreton, M., van Holst, R. J., & Luigjes, J. (2022). Metacognition and the effect of incentive motivation in two compulsive disorders: Gambling disorder and obsessive–compulsive disorder. *Psychiatry and clinical neurosciences*, *76*(9), 437–449.

Hoven, M., Lebreton, M., Engelmann, J. B., Denys, D., Luigjes, J., & van Holst, R. J. (2019). Abnormalities of confidence in psychiatry: An overview and future perspectives. *Translational psychiatry*, *9*(1), 268.

Hoven, M., Luigjes, J., Denys, D., Rouault, M., & van Holst, R. J. (2023). How do confidence and self-beliefs relate in psychopathology: A transdiagnostic approach. *Nature Mental Health*, 1–9.

Hoven, M., Rouault, M., van Holst, R., & Luigjes, J. (2022). Differences in metacognitive functioning between obsessive-compulsive disorder patients and highly compulsive individuals from the general population.

Howe, P. D., Vargas-Sáenz, A., Hulbert, C. A., & Boldero, J. M. (2019). Predictors of gambling and problem gambling in victoria, australia. *PloS one*, *14*(1), e0209277.

Hu, X., Zheng, J., Su, N., Fan, T., Yang, C., Yin, Y., Fleming, S. M., & Luo, L. (2021). A bayesian inference model for metamemory. *Psychological Review*, *128*(5), 824.

Huq, S., Garety, P. A., & Hemsley, D. R. (1988). Probabilistic judgements in deluded and non-deluded subjects. *The Quarterly Journal of Experimental Psychology Section A*, *40*(4), 801–812.

Huys, Q. J., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature neuroscience*, *19*(3), 404–413.

Huys, Q. J., Moutoussis, M., & Williams, J. (2011). Are computational models of any use to psychiatry? *Neural Networks*, *24*(6), 544–551.

Hyman, S. E. (2007). Can neuroscience be integrated into the dsm-v? *Nature Reviews Neuroscience*, *8*(9), 725–732.

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., Sanislow, C., & Wang, P. (2010). Research domain criteria (rdoc): Toward a new classification framework for research on mental disorders.

Ionescu, T. (2012). Exploring the nature of cognitive flexibility. *New ideas in psychology*, *30*(2), 190–200.

Izuma, K., & Adolphs, R. (2013). Social manipulation of preference in the human brain. *Neuron*, *78*(3), 563–573.

Janowsky, J. S., Shimamura, A. P., & Squire, L. R. (1989). Memory and metamemory: Comparisons between patients with frontal lobe lesions and amnesic patients. *Psychobiology*, *17*(1), 3–11.

Jauregui, P., Urbiola, I., & Estevez, A. (2016). Metacognition in pathological gambling and its relationship with anxious and depressive symptomatology. *Journal of Gambling Studies*, *32*, 675–688.

Jönsson, F. U., Olsson, H., & Olsson, M. J. (2005). Odor emotionality affects the confidence in odor naming. *Chemical senses*, *30*(1), 29–35.

Kantner, J., Solinger, L. A., Grybinas, D., & Dobbins, I. G. (2019). Confidence carryover during interleaved memory and perception judgments. *Memory & cognition*, *47*, 195–211.

Kao, Y.-C., Davis, E. S., & Gabrieli, J. D. (2005). Neural correlates of actual and predicted memory formation. *Nature neuroscience*, *8*(12), 1776–1783.

Kappes, A., Harvey, A. H., Lohrenz, T., Montague, P. R., & Sharot, T. (2020). Confirmation bias in the utilization of others' opinion strength. *Nature neuroscience*, *23*(1), 130–137.

Karadag, F., Oguzhanoglu, N., Ozdel, O., Atesci, F. C., & Amuk, T. (2005). Memory function in patients with obsessive compulsive disorder and the problem of confidence in their memories: A clinical study. *anxiety*, *6*, 8.

Kavli, H., & Berntsen, W. (2005). Gambling habits and gambling problems in the population. *Oslo: MMI Research*.

Kelly, S. P., & O'Connell, R. G. (2013). Internal and external influences on the rate of sensory evidence accumulation in the human brain. *Journal of Neuroscience*, *33*(50), 19434–19441.

Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature*, *455*(7210), 227–231.

Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annu. Rev. Psychol.*, *55*, 623–655.

Keyes, K., Eaton, N., Krueger, R., Skodol, A., Wall, M., Grant, B., Siever, L., & Hasin, D. (2013). Thought disorder in the meta-structure of psychopathology. *Psychological medicine*, *43*(8), 1673–1683.

Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice certainty is informed by both evidence and decision time. *Neuron*, *84*(6), 1329–1342.

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *science*, *324*(5928), 759–764.

Kim, B.-R., Liss, A., Rao, M., Singer, Z., & Compton, R. J. (2012). Social deviance activates the brain's error-monitoring system. *Cognitive, Affective, & Behavioral Neuroscience*, *12*, 65–73.

Kircher, T. T., Koch, K., Stottmeister, F., & Durst, V. (2007). Metacognition and reflexivity in patients with schizophrenia. *Psychopathology*, *40*(4), 254–260.

Koellinger, P., & Treffers, T. (2015). Joy leads to overconfidence, and a simple countermeasure. *PLoS One*, *10*(12), e0143263.

Komura, Y., Nikkuni, A., Hirashima, N., Uetake, T., & Miyamoto, A. (2013). Responses of pulvinar neurons reflect a subject's confidence in visual categorization. *Nature neuroscience*, *16*(6), 749–755.

Kopčanová, M., Ince, R. A., & Benwell, C. S. (2023). Two distinct stimulus-locked eeg signatures reliably encode domain-general confidence during decision formation. *bioRxiv*, 2023–04.

Koriat, A., et al. (2006). *Metacognition and consciousness*. Institute of Information Processing; Decision Making, University of Haifa.

Koriat, A. (2012). When are two heads better than one and why? *Science*, *336*(6079), 360–362.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological review*, *103*(3), 490.

Koriat, A., & Levy-Sadot, R. (2001). The combined contributions of the cue-familiarity and accessibility heuristics to feelings of knowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(1), 34.

Köther, U., Veckenstedt, R., Vitzthum, F., Roesch-Ely, D., Pfueller, U., Scheu, F., & Moritz, S. (2012). "don't give me that look"—overconfidence in false mental state perception in schizophrenia. *Psychiatry research*, *196*(1), 1–8.

Kotov, R., Gamez, W., Schmidt, F., & Watson, D. (2010). Linking "big" personality traits to anxiety, depressive, and substance use disorders: A meta-analysis. *Psychological bulletin*, *136*(5), 768.

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature neuroscience*, *13*(10), 1292–1298.

Kristiansen, S., & Jensen, S. M. (2011). Prevalence of gambling problems among adolescents in the nordic countries: An overview of national gambling surveys 1997–2009. *International Journal of Social Welfare*, *20*(1), 75–86.

Krumpe, T., Gerjets, P., Rosenstiel, W., & Spüler, M. (2019). Decision confidence: Eeg correlates of confidence in different phases of an old/new recognition task. *Brain-Computer Interfaces*, *6*(4), 162–177.

Kuhn, K. M., & Sniezek, J. A. (1996). Confidence and uncertainty in judgmental forecasting: Differential effects of scenario presentation. *Journal of Behavioral Decision Making*, *9*(4), 231–247.

Lak, A., Costa, G. M., Romberg, E., Koulakov, A. A., Mainen, Z. F., & 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., & Kepecs, A. (2017). Midbrain dopamine neurons signal belief in choice accuracy during a perceptual decision. *Current Biology*, *27*(6), 821–832.

Lakey, C. E., Goodie, A. S., & Campbell, W. K. (2007). Frequent card playing and pathological gambling: The utility of the georgia gambling task and iowa gambling task for predicting pathology. *Journal of Gambling Studies*, *23*, 285–297.

Lapate, R. C., Samaha, J., Rokers, B., Postle, B. R., & Davidson, R. J. (2020). Perceptual metacognition of human faces is causally supported by function of the lateral prefrontal cortex. *Communications biology*, *3*(1), 1–10.

Lazarov, A., Dar, R., Liberman, N., & Oded, Y. (2012). Obsessive-compulsive tendencies and undermined confidence are related to reliance on proxies for internal states in a false feedback paradigm. *Journal of behavior therapy and experimental psychiatry*, *43*(1), 556–564.

Le Berre, A.-P., Pinon, K., Vabret, F., Pitel, A.-L., Allain, P., Eustache, F., & Beaunieux, H. (2010). Study of metamemory in patients with chronic alcoholism using a feeling-of-knowing episodic memory task. *Alcoholism: Clinical and Experimental Research*, *34*(11), 1888–1898.

Lebreton, M., Abitbol, R., Daunizeau, J., & Pessiglione, M. (2015). Automatic integration of confidence in the brain valuation signal. *Nature neuroscience*, *18*(8), 1159–1167.

Ledgerwood, D. M., Dyshniku, F., McCarthy, J. E., Ostojic-Aitkens, D., Forfitt, J., & Rumble, S. C. (2020). Gambling-related cognitive distortions in residential treatment for gambling disorder. *Journal of Gambling Studies*, *36*, 669–683.

Lee, A. L., de Gardelle, V., & Mamassian, P. (2021). Global visual confidence. *Psychonomic Bulletin & Review*, *28*, 1233–1242.

Lee, D. G., Daunizeau, J., & Pezzulo, G. (2021). Evidence or confidence: What is really monitored during a decision? *bioRxiv*, 2021–04.

Lempert, K. M., Chen, Y. L., & Fleming, S. M. (2015). Relating pupil dilation and metacognitive confidence during auditory decision-making. *PLoS One*, *10*(5), e0126588.

Lépine, J.-P., Gasquet, I., Kovess, V., Arbabzadeh-Bouchez, S., Negre-Pages, L., Nachbaur, G., & Gaudin, A. (2005). Prevalence and comorbidity of psychiatric disorders in the french general population. *L'encephale*, *31*(2), 182–194.

Levy, D. J., & Glimcher, P. W. (2012). The root of all value: A neural common currency for choice. *Current opinion in neurobiology*, *22*(6), 1027–1038.

Li, E., Lavoie, S., Whitford, T., Moritz, S., Nelson, B., et al. (2018). Impaired action self-monitoring and cognitive confidence among ultra-high risk for psychosis and first-episode psychosis patients. *European Psychiatry*, *47*, 67–75.

Lin, Y., Gu, R., Luan, S., Hu, L., Qin, S., & Luo, Y.-j. (2021). The hierarchical sensitivity to social misalignment during decision-making under uncertainty. *Social Cognitive and Affective Neuroscience*, *16*(6), 565–575.

Logan, G. D., & Crump, M. J. (2010). Cognitive illusions of authorship reveal hierarchical error detection in skilled typists. *Science*, *330*(6004), 683–686.

Lorains, F. K., Cowlishaw, S., & Thomas, S. A. (2011). Prevalence of comorbid disorders in problem and pathological gambling: Systematic review and meta-analysis of population surveys. *Addiction*, *106*(3), 490–498.

Lundeberg, M. A., Fox, P. W., & Punćcohaŕ, J. (1994). Highly confident but wrong: Gender differences and similarities in confidence judgments. *Journal of educational psychology*, *86*(1), 114.

Lysaker, P. H., Minor, K. S., Lysaker, J. T., Hasson-Ohayon, I., Bonfils, K., Hochheiser, J., & Vohs, J. L. (2020). Metacognitive function and fragmentation in schizophrenia: Relationship to cognition, self-experience and developing treatments. *Schizophrenia Research: Cognition*, *19*, 100142.

Macdonald, P. A., Antony, M. M., Macleod, C. M., & Richter, M. A. (1997). Memory and confidence in memory judgments among individuals with obsessive compulsive disorder and non-clinical controls. *Behaviour research and therapy*, *35*(6), 497–505.

Mahmoodi, A., Bahrami, B., & Mehring, C. (2018). Reciprocity of social influence. *Nature communications*, *9*(1), 2474.

Mahmoodi, A., Bang, D., Olsen, K., Zhao, Y. A., Shi, Z., Broberg, K., Safavi, S., Han, S., Nili Ahmadabadi, M., Frith, C. D., et al. (2015). Equality bias impairs collective decision-making across cultures. *Proceedings of the National Academy of Sciences*, *112*(12), 3835–3840.

Mahmoodi, A., Ringwald, K., Wittmann, M. K., & Mehring, C. (2019). Social context alters metacognition.

Malenka, R. C., Angel, R. W., Hampton, B., & Berger, P. A. (1982). Impaired central error-correcting behavior in schizophrenia. *Archives of General Psychiatry*, *39*(1), 101–107.

Mallorquí-Bagué, N., Vintró-Alcaraz, C., Verdejo-García, A., Granero, R., Fernández-Aranda, F., Magaña, P., Mena-Moreno, T., Aymamı´, N., Gómez-Peña, M., Del Pino-Gutiérrez, A., et al. (2019). Impulsivity and cognitive distortions in different clinical phenotypes of gambling disorder: Profiles and longitudinal prediction of treatment outcomes. *European Psychiatry*, *61*, 9–16.

Mamassian, P., & de Gardelle, V. (2021). Modeling perceptual confidence and the confidence forced-choice paradigm. *Psychological Review*.

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and cognition*, *21*(1), 422–430.

Mann, L., Radford, M., Burnett, P., Ford, S., Bond, M., Leung, K., Nakamura, H., Vaughan, G., & Yang, K.-S. (1998). Cross-cultural differences in self-reported decision-making style and confidence. *International journal of psychology*, *33*(5), 325–335.

Marin, R. S., Biedrzycki, R. C., & Firinciogullari, S. (1991). Reliability and validity of the apathy evaluation scale. *Psychiatry research*, *38*(2), 143–162.

Masset, P., Ott, T., Lak, A., Hirokawa, J., & Kepecs, A. (2020). Behavior-and modality-general representation of confidence in orbitofrontal cortex. *Cell*, *182*(1), 112–126.

Massoni, S. (2014). Emotion as a boost to metacognition: How worry enhances the quality of confidence. *Consciousness and cognition*, *29*, 189–198.

McCurdy, L. Y., Maniscalco, B., Metcalfe, J., Liu, K. Y., De Lange, F. P., & Lau, H. (2013). Anatomical coupling between distinct metacognitive systems for memory and visual perception. *Journal of Neuroscience*, *33*(5), 1897–1906.

McNally, R. J., & Kohlbeck, P. A. (1993). Reality monitoring in obsessive-compulsive disorder. *Behaviour research and therapy*, *31*(3), 249–253.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, *27*(1), 415–444.

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review*, *15*(1), 174–179.

Meyer, G., Hayer, T., & Griffiths, M. (2009). *Problem gambling in europe: Challenges, prevention, and interventions*. Springer Science & Business Media.

Meyniel, F., Sigman, M., & Mainen, Z. F. (2015). Confidence as bayesian probability: From neural origins to behavior. *Neuron*, *88*(1), 78–92.

Middlebrooks, P. G., & Sommer, M. A. (2012). Neuronal correlates of metacognition in primate frontal cortex. *Neuron*, *75*(3), 517–530.

Mintzer, M. Z., & Stitzer, M. L. (2002). Cognitive impairment in methadone maintenance patients. *Drug and alcohol dependence*, *67*(1), 41–51.

Miyamoto, K., Osada, T., Setsuie, R., Takeda, M., Tamura, K., Adachi, Y., & Miyashita, Y. (2017). Causal neural network of metamemory for retrospection in primates. *Science*, *355*(6321), 188–193.

Moeller, S. J., Fleming, S. M., Gan, G., Zilverstand, A., Malaker, P., Schneider, K. E., Preston-Campbell, R. N., Parvaz, M. A., Maloney, T., Alia-Klein, N., et al. (2016). Metacognitive impairment in active cocaine use disorder is associated with individual differences in brain structure. *European Neuropsychopharmacology*, *26*(4), 653–662.

Morales, J., Lau, H., & Fleming, S. M. (2018). Domain-general and domain-specific patterns of activity supporting metacognition in human prefrontal cortex. *Journal of Neuroscience*, *38*(14), 3534–3546.

Moritz, S., Jacobsen, D., Willenborg, B., Jelinek, L., & Fricke, S. (2006). A check on the memory deficit hypothesis of obsessive–compulsive checking. *European archives of psychiatry and clinical neuroscience*, *256*, 82–86.

Moritz, S., & Jaeger, A. (2018). Decreased memory confidence in obsessive–compulsive disorder for scenarios high and low on responsibility: Is low still too high? *European archives of psychiatry and clinical neuroscience*, *268*, 291–299.

Moritz, S., & Woodward, T. S. (2002). Memory confidence and false memories in schizophrenia. *The Journal of nervous and mental disease*, *190*(9), 641–643.

Moritz, S., Woodward, T. S., Cuttler, C., Whitman, J. C., & Watson, J. M. (2004). False memories in schizophrenia. *Neuropsychology*, *18*(2), 276.

Moritz, S., Woodward, T. S., Whitman, J. C., & Cuttler, C. (2005). Confidence in errors as a possible basis for delusions in schizophrenia. *The Journal of nervous and mental disease*, *193*(1), 9–16.

Moritz, S., Woodward, T., & Ruff, C. (2003). Source monitoring and memory confidence in schizophrenia. *Psychological medicine*, *33*(1), 131–139.

Moritz, S., Woznica, A., Andreou, C., & Köther, U. (2012). Response confidence for emotion perception in schizophrenia using a continuous facial sequence task. *Psychiatry Research*, *200*(2-3), 202–207.

Moses-Payne, M. E., Rollwage, M., Fleming, S. M., & Roiser, J. P. (2019). Postdecision evidence integration and depressive symptoms. *Frontiers in psychiatry*, *10*, 639.

Murphy, P. R., Robertson, I. H., Harty, S., & O'Connell, R. G. (2015). Neural evidence accumulation persists after choice to inform metacognitive judgments. *Elife*, *4*, e11946.

Nelson, T. O. (1987). The goodman-kruskal gamma coefficient as an alternative to signal-detection theory's measures of absolute-judgment accuracy.

Nelson, T. O., Dunlosky, J., Graf, A., & Narens, L. (1994). Utilization of metacognitive judgments in the allocation of study during multitrial learning. *Psychological Science*, *5*(4), 207–213.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In *Psychology of learning and motivation* (pp. 125–173). Elsevier.

Newson, J. J., Pastukh, V., & Thiagarajan, T. C. (2021). Poor separation of clinical symptom profiles by dsm-5 disorder criteria. *Frontiers in Psychiatry*, *12*, 775762.

Nicolò, G., Dimaggio, G., Popolo, R., Carcione, A., Procacci, M., Hamm, J., Buck, K. D., Pompili, E., Buccione, I., Lagrotteria, B., et al. (2012). Associations of metacognition with symptoms, insight, and neurocognition in clinically stable outpatients with schizophrenia. *The Journal of nervous and mental disease*, *200*(7), 644–647.

O'connell, R. G., Dockree, P. M., & Kelly, S. P. (2012). A supramodal accumulation-to-bound signal that determines perceptual decisions in humans. *Nature neuroscience*, *15*(12), 1729–1735.

O'Kearney, E., Brown, C., Jelinek, G., Neate, S., Taylor, K., Bevens, W., De Livera, A., Simpson Jr, S., & Weiland, T. (2020). Mastery is associated with greater physical and mental health-related quality of life in two international cohorts of people with multiple sclerosis. *Multiple sclerosis and related disorders*, *38*, 101481.

Olawole-Scott, H., & Yon, D. (2023). Expectations about precision bias metacognition and awareness. *Journal of Experimental Psychology: General*.

Orgaz, C., Estévez, A., & Matute, H. (2013). Pathological gamblers are more vulnerable to the illusion of control in a standard associative learning task. *Frontiers in psychology*, *4*, 306.

Pallier, G., Wilkinson, R., Danthiir, V., Kleitman, S., Knezevic, G., Stankov, L., & Roberts, R. D. (2002). The role of individual differences in the accuracy of confidence judgments. *The Journal of general psychology*, *129*(3), 257–299.

Pannu, J. K., Kaszniak, A. W., & Rapcsak, S. Z. (2005). Metamemory for faces following frontal lobe damage. *Journal of the International Neuropsychological Society*, *11*(6), 668–676.

Parra, L., Alvino, C., Tang, A., Pearlmutter, B., Yeung, N., Osman, A., & Sajda, P. (2002). Linear spatial integration for single-trial detection in encephalography. *NeuroImage*, *17*(1), 223–230.

Patton, J. H., Stanford, M. S., & Barratt, E. S. (1995). Factor structure of the barratt impulsiveness scale. *Journal of clinical psychology*, *51*(6), 768–774.

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

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

Penrod, S., & Cutler, B. (1995). Witness confidence and witness accuracy: Assessing their forensic relation. *Psychology, Public Policy, and Law*, *1*(4), 817.

Perandrés-Gómez, A., Navas, J. F., van Timmeren, T., & Perales, J. C. (2021). Decision-making (in) flexibility in gambling disorder. *Addictive behaviors*, *112*, 106534.

Pescetelli, N., Rees, G., & Bahrami, B. (2016). The perceptual and social components of metacognition. *Journal of Experimental Psychology: General*, *145*(8), 949.

Pescetelli, N., Hauperich, A.-K., & Yeung, N. (2021). Confidence, advice seeking and changes of mind in decision making. *Cognition*, *215*, 104810.

Pescetelli, N., & Yeung, N. (2020). The effects of recursive communication dynamics on belief updating. *Proceedings of the Royal Society B*, *287*(1931), 20200025.

Pescetelli, N., & Yeung, N. (2022). Benefits of spontaneous confidence alignment between dyad members. *Collective Intelligence*, *1*(2), 26339137221126915.

Peters, M. J., Hauschildt, M., Moritz, S., & Jelinek, L. (2013). Impact of emotionality on memory and meta-memory in schizophrenia using video sequences. *Journal of Behavior Therapy and Experimental Psychiatry*, *44*(1), 77–83.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, *27*(2), 169–190.

Pisauro, M. A., Fouragnan, E., Retzler, C., & Philiastides, M. G. (2017). Neural correlates of evidence accumulation during value-based decisions revealed via simultaneous eeg-fmri. *Nature communications*, *8*(1), 15808.

Polanı´a, R., Moisa, M., Opitz, A., Grueschow, M., & Ruff, C. C. (2015). The precision of value-based choices depends causally on fronto-parietal phase coupling. *Nature communications*, *6*(1), 8090.

Polich, J., & Kok, A. (1995). Cognitive and biological determinants of p300: An integrative review. *Biological psychology*, *41*(2), 103–146.

Portin, R., Kovala, T., Polo-Kantola, P., Revonsuo, A., Müller, K., & Matikainen, E. (2000). Does p3 reflect attentional or memory performances, or cognition more generally? *Scandinavian Journal of Psychology*, *41*(1), 31–40.

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: Distinct probabilistic quantities for different goals. *Nature neuroscience*, *19*(3), 366–374.

Price, P. C., & Stone, E. R. (2004). Intuitive evaluation of likelihood judgment producers: Evidence for a confidence heuristic. *Journal of Behavioral Decision Making*, *17*(1), 39–57.

Proust, J. (2010). Metacognition. *Philosophy Compass*, *5*(11), 989–998.

Pulford, B. D., Colman, A. M., Buabang, E. K., & Krockow, E. M. (2018). The persuasive power of knowledge: Testing the confidence heuristic. *Journal of Experimental Psychology: General*, *147*(10), 1431.

Qiu, L., Su, J., Ni, Y., Bai, Y., Zhang, X., Li, X., & Wan, X. (2018). The neural system of metacognition accompanying decision-making in the prefrontal cortex. *PLoS biology*, *16*(4), e2004037.

Quiles, C., Prouteau, A., & Verdoux, H. (2015). Associations between self-esteem, anxiety and depression and metacognitive awareness or metacognitive knowledge. *Psychiatry research*, *230*(2), 738–741.

Rabbitt, P., & Rodgers, B. (1977). What does a man do after he makes an error? an analysis of response programming. *Quarterly Journal of Experimental Psychology*, *29*(4), 727–743.

Rahnev, D. (2023). Measuring metacognition: A comprehensive assessment of current methods.

Rahnev, D., Desender, K., Lee, A. L., Adler, W. T., Aguilar-Lleyda, D., Akdoğan, B., Arbuzova, P., Atlas, L. Y., Balcı, F., Bang, J. W., et al. (2020). The confidence database. *Nature human behaviour*, *4*(3), 317–325.

Rahnev, D., Koizumi, A., McCurdy, L. Y., D'Esposito, M., & Lau, H. (2015). Confidence leak in perceptual decision making. *Psychological science*, *26*(11), 1664–1680.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological science*, *9*(5), 347–356.

Ratcliffe, M., & Wilkinson, S. (2015). Thought insertion clarified. *Journal of Consciousness Studies*, *22*(11-12), 246–269.

Rausch, M., Zehetleitner, M., Steinhauser, M., & Maier, M. E. (2020). Cognitive modelling reveals distinct electrophysiological markers of decision confidence and error monitoring. *NeuroImage*, *218*, 116963.

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in cognitive sciences*, *20*(9), 676–688.

Rivers, M. L., Fitzsimmons, C. J., Fisk, S. R., Dunlosky, J., & Thompson, C. A. (2021). Gender differences in confidence during number-line estimation. *Metacognition and Learning*, *16*, 157–178.

Rosenberg, M. (1965). Rosenberg self-esteem scale. *Journal of Religion and Health*.

Rouault, M., Dayan, P., & Fleming, S. M. (2019). Forming global estimates of self-performance from local confidence. *Nature communications*, *10*(1), 1141.

Rouault, M., & Fleming, S. M. (2020). Formation of global self-beliefs in the human brain. *Proceedings of the National Academy of Sciences*, *117*(44), 27268–27276.

Rouault, M., McWilliams, A., Allen, M. G., & Fleming, S. M. (2018). Human metacognition across domains: Insights from individual differences and neuroimaging. *Personality neuroscience*, *1*.

Rouault, M., Seow, T., Gillan, C. M., & Fleming, S. M. (2018). Psychiatric symptom dimensions are associated with dissociable shifts in metacognition but not task performance. *Biological psychiatry*, *84*(6), 443–451.

Rouault, M., Will, G.-J., Fleming, S. M., & Dolan, R. J. (2022). Low self-esteem and the formation of global self-performance estimates in emerging adulthood. *Translational Psychiatry*, *12*(1), 272.

Rousseau, L., & Kashur, N. (2021). Socially shared feelings of imminent recall: More tip-of-the-tongue states are experienced in small groups. *Frontiers in Psychology*, *12*, 704433.

Sadeghi, S., Ekhtiari, H., Bahrami, B., & Ahmadabadi, M. N. (2017). Metacognitive deficiency in a perceptual but not a memory task in methadone maintenance patients. *Scientific reports*, *7*(1), 1–8.

Sadras, N., Sani, O. G., Ahmadipour, P., & Shanechi, M. M. (2022). Post-stimulus encoding of decision confidence in eeg: Toward a brain-computer interface for decision making. *bioRxiv*, 2022–11.

Salti, M., Bar-Haim, Y., & Lamy, D. (2012). The p3 component of the erp reflects conscious perception, not confidence. *Consciousness and cognition*, *21*(2), 961–968.

Samaha, J., Iemi, L., & Postle, B. R. (2017). Prestimulus alpha-band power biases visual discrimination confidence, but not accuracy. *Consciousness and cognition*, *54*, 47–55.

Sauer, J. D., & Brewer, N. (2015). Confidence and accuracy of eyewitness identification. *Forensic facial identification: Theory and practice of identification from eyewitnesses, composites and CCTV*, 185–208.

Schaefer, P. S., Williams, C. C., Goodie, A. S., & Campbell, W. K. (2004). Overconfidence and the big five. *Journal of research in Personality*, *38*(5), 473–480.

Schevernels, H., Krebs, R. M., Santens, P., Woldorff, M. G., & Boehler, C. N. (2014). Task preparation processes related to reward prediction precede those related to task-difficulty expectation. *NeuroImage*, *84*, 639–647.

Schnyer, D. M., Verfaellie, M., Alexander, M. P., LaFleche, G., Nicholls, L., & Kaszniak, A. W. (2004). A role for right medial prefrontal cortex in accurate feeling-of-knowing judgments: Evidence from patients with lesions to frontal cortex. *Neuropsychologia*, *42*(7), 957–966.

Schraw, G. (1996). The effect of generalized metacognitive knowledge on test performance and confidence judgments. *The Journal of Experimental Education*, *65*(2), 135–146.

Schwartz, B. L., & Díaz, F. (2014). Quantifying human metacognition for the neurosciences. In *The cognitive neuroscience of metacognition* (pp. 9–23). Springer.

Selimbeyoglu, A., Keskin-Ergen, Y., & Demiralp, T. (2012). What if you are not sure? electroencephalographic correlates of subjective confidence level about a decision. *Clinical Neurophysiology*, *123*(6), 1158–1167.

Seow, T. X., & Gillan, C. M. (2020). Transdiagnostic phenotyping reveals a host of metacognitive deficits implicated in compulsivity. *Scientific reports*, *10*(1), 1–11.

Seow, T. X., Rouault, M., Gillan, C. M., & Fleming, S. M. (2021). How local and global metacognition shape mental health. *Biological psychiatry*, *90*(7), 436–446.

Shachar, A. B., Lazarov, A., Goldsmith, M., Moran, R., & Dar, R. (2013). Exploring metacognitive components of confidence and control in individuals with obsessive-compulsive tendencies. *Journal of behavior therapy and experimental psychiatry*, *44*(2), 255–261.

Sharma, V., Sagar, R., Kaloiya, G., & Mehta, M. (2022). The scope of metacognitive therapy in the treatment of psychiatric disorders. *Cureus*, *14*(3).

Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Supra-personal cognitive control and metacognition. *Trends in cognitive sciences*, *18*(4), 186–193.

Shea, N., & Frith, C. D. (2016). Dual-process theories and consciousness: The case for 'type zero'cognition. *Neuroscience of Consciousness*, *2016*(1).

Shea, N., & Frith, C. D. (2019). The global workspace needs metacognition. *Trends in cognitive sciences*, *23*(7), 560–571.

Shekhar, M., & Rahnev, D. (2018). Distinguishing the roles of dorsolateral and anterior pfc in visual metacognition. *Journal of Neuroscience*, *38*(22), 5078–5087.

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217–240.

Shenhav, A., Prater Fahey, M., & Grahek, I. (2021). Decomposing the motivation to exert mental effort. *Current Directions in Psychological Science*, *30*(4), 307–314.

Sherif, M. (1935). A study of some social factors in perception. *Archives of Psychology (Columbia University)*.

Shestakova, A., Rieskamp, J., Tugin, S., Ossadtchi, A., Krutitskaya, J., & Klucharev, V. (2013). Electrophysiological precursors of social conformity. *Social cognitive and affective neuroscience*, *8*(7), 756–763.

Shimamura, A. P., & Squire, L. R. (1986). Memory and metamemory: A study of the feeling-of-knowing phenomenon in amnesic patients. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*(3), 452.

Shooshtari, S. V., Sadrabadi, J. E., Azizi, Z., & Ebrahimpour, R. (2019). Confidence representation of perceptual decision by eeg and eye data in a random dot motion task. *Neuroscience*, *406*, 510–527.

Silver, I., Mellers, B. A., & Tetlock, P. E. (2021). Wise teamwork: Collective confidence calibration predicts the effectiveness of group discussion. *Journal of Experimental Social Psychology*, *96*, 104157.

Silverstone, P. H., & Salsali, M. (2003). Low self-esteem and psychiatric patients: Part i–the relationship between low self-esteem and psychiatric diagnosis. *Annals of general hospital psychiatry*, *2*(1), 1–9.

Sims, C. A. (2003). Implications of rational inattention. *Journal of monetary Economics*, *50*(3), 665–690.

Smith, J. L., Johnstone, S. J., & Barry, R. J. (2008). Movement-related potentials in the go/nogo task: The p3 reflects both cognitive and motor inhibition. *Clinical neurophysiology*, *119*(3), 704–714.

Sniezek, J. A., & Buckley, T. (1991). Confidence depends on level of aggregation. *Journal of Behavioral Decision Making*, *4*(4), 263–272.

Sniezek, J. A., & Van Swol, L. M. (2001). Trust, confidence, and expertise in a judge-advisor system. *Organizational behavior and human decision processes*, *84*(2), 288–307.

Sowislo, J. F., & Orth, U. (2013). Does low self-esteem predict depression and anxiety? a meta-analysis of longitudinal studies. *Psychological bulletin*, *139*(1), 213.

Speilberger, C., Gorsuch, R., Lushene, R., Vagg, R., & Jacobs, G. (1983). Manual for the stait-trait anxiety inventory.

Squires, K. C., Squires, N. K., & Hillyard, S. A. (1975a). Decision-related cortical potentials during an auditory signal detection task with cued observation intervals. *Journal of Experimental Psychology: Human Perception and Performance*, *1*(3), 268.

Squires, K. C., Squires, N. K., & Hillyard, S. A. (1975b). Vertex evoked potentials in a rating-scale detection task: Relation to signal probability. *Behavioral biology*, *13*(1), 21–34.

Steinhauser, M., & Yeung, N. (2010). Decision processes in human performance monitoring. *Journal of Neuroscience*, *30*(46), 15643–15653.

Stephan, K. E., & Mathys, C. (2014). Computational approaches to psychiatry. *Current opinion in neurobiology*, *25*, 85–92.

Stone, E. R., Dodrill, C. L., & Johnson, N. (2001). Depressive cognition: A test of depressive realism versus negativity using general knowledge questions. *The Journal of Psychology*, *135*(6), 583–602.

Subramaniam, M., Wang, P., Soh, P., Vaingankar, J. A., Chong, S. A., Browning, C. J., & Thomas, S. A. (2015). Prevalence and determinants of gambling disorder among older adults: A systematic review. *Addictive Behaviors*, *41*, 199–209.

Sun, X., Zhu, C., & So, S. (2017). Dysfunctional metacognition across psychopathologies: A meta-analytic review. *European Psychiatry*, *45*, 139–153.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Sutton, S., Braren, M., Zubin, J., & John, E. (1965). Evoked-potential correlates of stimulus uncertainty. *Science*, *150*(3700), 1187–1188.

Sutton, S., Ruchkin, D. S., Munson, R., Kietzman, M. L., & Hammer, M. (1982). Event-related potentials in a two-interval forced-choice detection task. *Perception & Psychophysics*, *32*, 360–374.

Tagliabue, C. F., Veniero, D., Benwell, C. S., Cecere, R., Savazzi, S., & Thut, G. (2019). The eeg signature of sensory evidence accumulation during decision

formation closely tracks subjective perceptual experience. *Scientific reports*, *9*(1), 1–12.

Tenney, E. R., MacCoun, R. J., Spellman, B. A., & Hastie, R. (2007). Calibration trumps confidence as a basis for witness credibility. *Psychological Science*, *18*(1), 46–50.

Terrace, H. S., & Son, L. K. (2009). Comparative metacognition. *Current opinion in neurobiology*, *19*(1), 67–74.

Thomas, J. P., & McFadyen, R. G. (1995). The confidence heuristic: A game-theoretic analysis. *Journal of Economic Psychology*, *16*(1), 97–113.

Tolin, D. F., Abramowitz, J. S., Brigidi, B. D., Amir, N., Street, G. P., & Foa, E. B. (2001). Memory and memory confidence in obsessive–compulsive disorder. *Behaviour research and therapy*, *39*(8), 913–927.

Treadwell, J. R., & Nelson, T. O. (1996). Availability of information and the aggregation of confidence in prior decisions. *Organizational Behavior and Human Decision Processes*, *68*(1), 13–27.

Trouche, E., Sander, E., & Mercier, H. (2014). Arguments, more than confidence, explain the good performance of reasoning groups. *Journal of Experimental Psychology: General*, *143*(5), 1958.

Tsuang, M. T., Lyons, M. J., & Faraone, S. V. (1990). Heterogeneity of schizophrenia. *The British Journal of Psychiatry*, *156*(1), 17–26.

Tugin, S., Gorin, A., Kanunikov, I., & Shestakova, A. (2015). Hyperscanning of social attunement: An frn study. *Psychology. Journal of Higher School of Economics*, *12*(4), 48–63.

Tükel, R., Polat, A., Özdemir, Ö., Aksüt, D., & Türksoy, N. (2002). Comorbid conditions in obsessive-compulsive disorder. *Comprehensive psychiatry*, *43*(3), 204–209.

Uddin, L. Q. (2021). Cognitive and behavioural flexibility: Neural mechanisms and clinical considerations. *Nature Reviews Neuroscience*, *22*(3), 167–179.

Vaccaro, A. G., & Fleming, S. M. (2018). Thinking about thinking: A coordinate-based meta-analysis of neuroimaging studies of metacognitive judgements. *Brain and neuroscience advances*, *2*, 2398212818810591.

Valeriani, D., Cinel, C., & Poli, R. (2017). Group augmentation in realistic visual-search decisions via a hybrid brain-computer interface. *Scientific reports*, *7*(1), 7772.

Valeriani, D., O'Flynn, L. C., Worthley, A., Sichani, A. H., & Simonyan, K. (2022). Multimodal collaborative brain-computer interfaces aid human-machine team decision-making in a pandemic scenario. *Journal of Neural Engineering*, *19*(5), 056036.

Valk, S. L., Bernhardt, B. C., Böckler, A., Kanske, P., & Singer, T. (2016). Substrates of metacognition on perception and metacognition on higher-order cognition relate to different subsystems of the mentalizing network. *Human Brain Mapping*, *37*(10), 3388–3399.

Van den Berg, R., Zylberberg, A., Kiani, R., Shadlen, M. N., & Wolpert, D. M. (2016). Confidence is the bridge between multi-stage decisions. *Current Biology*, *26*(23), 3157–3168.

van den Hout, M., & Kindt, M. (2003). Repeated checking causes memory distrust. *Behaviour research and therapy*, *41*(3), 301–316.

van der Wel, P., & Van Steenbergen, H. (2018). Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic bulletin & review*, *25*, 2005–2015.

Van Marcke, H., Le Denmat, P., Verguts, T., & Desender, K. (2022). Manipulating prior beliefs causally induces under-and overconfidence. *BioRxiv*, 2022–03.

Van Swol, L., & Sniezek, J. (2002). Trust me, i'm an expert: Trust and confidence and acceptance of expert advice. *8th International Conference on Behavioral Decision Research in Management (BDRAM), Chicago, Illnois, USA*.

Van Timmeren, T., Daams, J. G., Van Holst, R. J., & Goudriaan, A. E. (2018). Compulsivity-related neurocognitive performance deficits in gambling disorder: A systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews*, *84*, 204–217.

Verdejo-García, A., Lawrence, A. J., & Clark, L. (2008). Impulsivity as a vulnerability marker for substance-use disorders: Review of findings from high-risk research, problem gamblers and genetic association studies. *Neuroscience & Biobehavioral Reviews*, *32*(4), 777–810.

Volberg, R. A., & Bernhard, B. (2006). *The 2006 study of gambling and problem gambling in new mexico: Report to the responsible gaming association of new mexico* (tech. rep.). Responsible Gaming Association of New Mexico.

Wardle, H. (2007). *British gambling prevalence survey 2007*. The Stationery Office.

Warman, D. M. (2008). Reasoning and delusion proneness: Confidence in decisions. *The Journal of nervous and mental disease*, *196*(1), 9–15.

Wells, A. (2011). *Metacognitive therapy for anxiety and depression*. Guilford press.

Welte, J. W., Barnes, G. M., Tidwell, M.-C. O., & Wieczorek, W. F. (2017). Predictors of problem gambling in the us. *Journal of Gambling Studies*, *33*, 327–342.

Wilkinson, R. T., & Seales, D. M. (1978). Eeg event-related potentials and signal detection. *Biological psychology*, *7*(1-2), 13–28.

Will, G.-J., Rutledge, R. B., Moutoussis, M., & Dolan, R. J. (2017). Neural and computational processes underlying dynamic changes in self-esteem. *Elife*, *6*, e28098.

Winne, P. H. (2017). Cognition and metacognition within self-regulated learning. In *Handbook of self-regulation of learning and performance* (pp. 36–48). Routledge.

Winne, P. H., & Perry, N. E. (2000). Measuring self-regulated learning. In *Handbook of self-regulation* (pp. 531–566). Elsevier.

Wise, T., & Dolan, R. J. (2020). Associations between aversive learning processes and transdiagnostic psychiatric symptoms in a general population sample. *Nature communications*, *11*(1), 4179.

Wise, T., Robinson, O. J., & Gillan, C. M. (2023). Identifying transdiagnostic mechanisms in mental health using computational factor modeling. *Biological Psychiatry*, *93*(8), 690–703.

Wittmann, M. K., Kolling, N., Faber, N. S., Scholl, J., Nelissen, N., & Rushworth, M. F. (2016). Self-other mergence in the frontal cortex during cooperation and competition. *Neuron*, *91*(2), 482–493.

Wyckmans, F., Otto, A. R., Sebold, M., Daw, N., Bechara, A., Saeremans, M., Kornreich, C., Chatard, A., Jaafari, N., & Noël, X. (2019). Reduced model-based decision-making in gambling disorder. *Scientific reports*, *9*(1), 19625.

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational behavior and human decision processes*, *83*(2), 260–281.

Yates, J. F., Price, P. C., Lee, J.-W., & Ramirez, J. (1996). Good probabilistic forecasters: The 'consumer's' perspective. *International Journal of Forecasting*, *12*(1), 41–56.

Yeung, N. (2014). Conflict monitoring and cognitive control. *Ochsner KN, Kosslyn SM (eds) The Oxford Handbook of Cognitive Neuroscience, Volume 2: The Cutting Edges*, 275–299.

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: Confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1310–1321.

Yılmaz, A. E., Gençöz, T., & Wells, A. (2011). The temporal precedence of metacognition in the development of anxiety and depression symptoms in the context of life-stress: A prospective study. *Journal of Anxiety Disorders*, *25*(3), 389–396.

Young, M., & Stevens, M. (2008). Sogs and cgpi: Parallel comparison on a diverse population. *Journal of Gambling Studies*, *24*, 337–356.

Yu, R., & Sun, S. (2013). To conform or not to conform: Spontaneous conformity diminishes the sensitivity to monetary outcomes. *PloS one*, *8*(5), e64530.

Zakrzewski, A. C., Wisniewski, M. G., Iyer, N., & Simpson, B. D. (2019). Confidence tracks sensory-and decision-related erp dynamics during auditory detection. *Brain and Cognition*, *129*, 49–58.

Zarnoth, P., & Sniezek, J. A. (1997). The social influence of confidence in group decision making. *Journal of Experimental Social Psychology*, *33*(4), 345–366.

Zhang, Z., Wang, M., Miao, X., Li, Y., Hitchman, G., & Yuan, Z. (2017). Individuals with high obsessive-compulsive tendencies or undermined confidence rely more on external proxies to access their internal states. *Journal of Behavior Therapy and Experimental Psychiatry*, *54*, 263–269.

Zimmerman, B. J. (1986). Becoming a self-regulated learner: Which are the key subprocesses? *Contemporary educational psychology*, *11*(4), 307–313.

Zung, W. W. (1965). A self-rating depression scale. *Archives of general psychiatry*, *12*(1), 63–70.