Ch. 32, Vol. 2

Cognitive Neuroscience of Metacognition

Maja Friedemann, Dan Bang, and Nick Yeung

Introduction

This chapter explores the neural underpinnings of metacognition – the monitoring and controlling of one's own thoughts. After first defining metacognition and describing how it can be measured, we discuss the evidence for explicit representations of metacognitive signals, reviewing neuroimaging studies that have dissociated metacognitive from object-level processes. Further, we consider the functional value of metacognition, including the social sharing of confidence estimates for group decision-making. Finally, we review research debates that cognitive neuroscience research may help address, focusing on how metacognition is represented in the brain, the relationship between metacognition and Theory of Mind, and the nature of metacognitive deficits in psychiatric disorders.

What is metacognition?

Metacognition, first conceived by Flavell (1979), is *"knowledge and cognition about cognitive phenomena"*. Simply put, metacognition is thinking about thinking. 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. Similarly, choosing what lecture content to review employs metacognition as it depends on evaluating your own level of understanding of the lecture material. Metacognition thus 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, in which subjective confidence estimates can have dramatic consequences. Metacognition initially received most attention in applied fields such as educational and organisational psychology (Bonaccio and Dalal, 2006; Sniezek and Van Swol, 2001; Zimmerman, 1986), but more recently also has been investigated through the lens of cognitive neuroscience (Fleming and Dolan, 2012; Shimamura and Squire, 1986; Vaccaro and Fleming, 2018). Not only can this research tell us something about the neurobiological basis, but it may also help elucidate underlying mechanisms and inform psychological theories of metacognition.

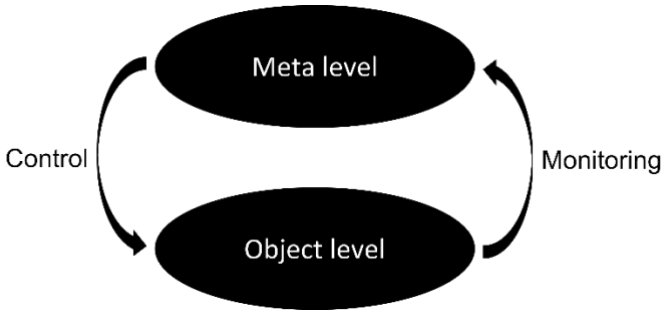

Figure 1. Model of metacognition (adapted from Nelson and Narens, 1990)

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 (Figure 1). As this hierarchical layering of cognition may implicate a certain degree of self-awareness, metacognition has also been associated with consciousness (Koriat, 2007). The present chapter considers the cognitive neuroscience of these explicit evaluations of cognitive processes, for example as they are reflected in people's verbal reports or numerical ratings of the quality of their learning or the accuracy of their decisions.

Our review is necessarily selective. Although of interest, we do not discuss evidence that some metacognitive processes may also operate implicitly without conscious awareness (Shea and Frith, 2016), such that individuals respond more cautiously following errors even when they are unaware that they made any mistake (Logan and Crump, 2010). Such non-verbal expressions of metacognitive processes allow for studying metacognition also in non-human animals (Terrace and Son, 2009) and pre-verbal infants (Goupil and Kouider, 2019). Likewise we do not review research at the other end of the experiential spectrum, involving the subjective correlates of metacognition including, tip-of-the-tongue (TOT) experiences characterised by a powerful and tantalising sense that one knows something that cannot be recalled momentarily, "aha-moments" that reflect a sudden insight that one has understood or solved a problem, and déjà vu experiences that reflect a metacognitive realisation that a seemingly familiar scene has in fact not been perceived before (Metcalfe and Schwartz, 2016).

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 (FOK) (Reder and Ritter, 1992) and Judgements of Learning (JOL) (Rhodes, 2015; Son and Metcalfe, 2005). FOK ratings reflect participants' subjective feelings about whether they will later be able to recognise the correct answer to a question or a problem from a list of alternatives. JOL 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).

Quantitative measures of metacognition probe the association between subjective metacognitive judgements and objective task performance. Metacognition is typically assessed using measures such as metacognitive sensitivity, metacognitive bias, and metacognitive efficiency (Fleming and Lau, 2014). Metacognitive sensitivity (also called resolution) quantifies the ability to discriminate correct and incorrect judgements. For instance, someone with good metacognitive sensitivity would be expected to be objectively correct whenever they indicate high confidence or a high FOK. Metacognitive bias, also called calibration, evaluates whether someone is generally over- or underconfident relative to their objective accuracy (Figure 2). Finally, metacognitive efficiency quantifies the level of metacognitive sensitivity considering task performance, as the same person may display greater metacognitive sensitivity in a simple task compared to a difficult one or vice versa.

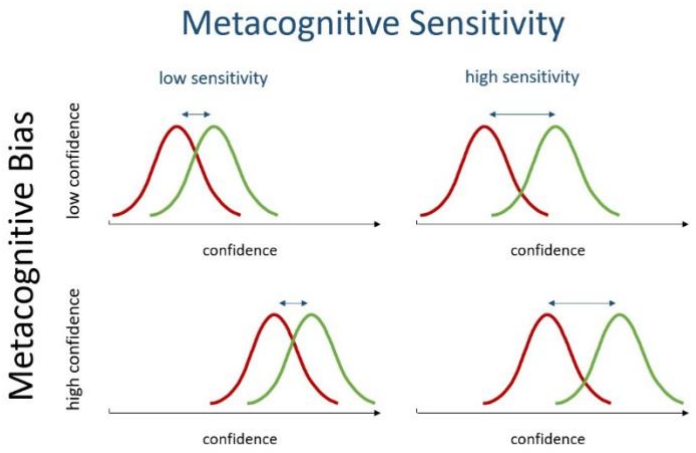

Figure 2. Distribution of confidence ratings for correct (green) and incorrect (red) choices Metacognitive bias refers to the overall level of confidence expressed. Distributions on the top have overall lower reported confidence level, while distributions on the bottom have an overall higher reported confidence level. Metacognitive sensitivity refers to the separation between

confidence distributions for correct and incorrect choices. Distributions on the left are closer together and do not separate well between correct and incorrect choices, whereas distributions on the right clearly distinguish between correct and incorrect choices.

Metacognitive sensitivity can be 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 commonly used correlational estimation of metacognitive sensitivity is via the Goodman-Kruskal gamma coefficient G (Nelson, 1984). 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 $accuracy_a < accuracy_b$; see Table 1).

Table 1. Illustrative frequency table displaying number correct and incorrect responses across five confidence bins

Reported confidence	0-19	20-39	40-59	60-79	80-100
N correct responses	0	4	6	6	11
N incorrect responses	6	3	2	2	0

Notes. Number of concordant pairs C = 6*(4+6+6+11) + 3*(6+6+11) + 2*(6+11) + 2*(11). Number of discordant pairs D = 3*(0) + 2*(0+4) + 2*(0+4+6) + 0*(0+4+6+6). Gamma (G) = (C-D) / (C+D), (i.e., 0.82 here), indicating a relatively strong relationship between subjective confidence and objective accuracy.

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 (Figure 3). Meta-d'

evaluates the type 1 d' (the sensitivity for discriminating stimulus alternatives (i.e., 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). It also allows for assessment of metacognitive efficiency in terms of the ratio between meta-d' and type 1 d'.

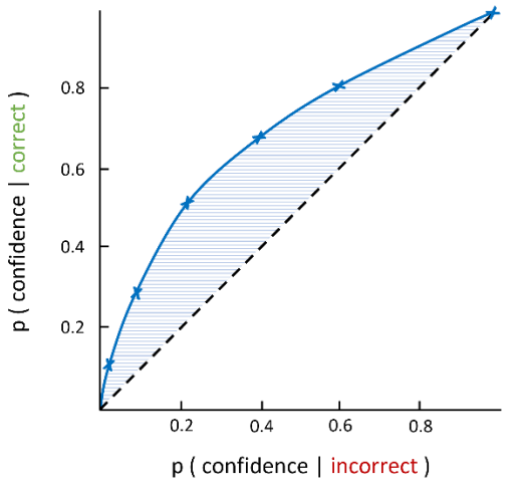

Figure 3. Illustrative Receiver Operating Characteristic (ROC) curve

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.

Neural substrates of metacognitive monitoring

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 FOKs 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. We next review key themes and debates emerging from this research, first regarding the distinct regions associated with metacognitive versus object-level processes, and then regarding suggested functional dissociations among different aspects of metacognition.

Metacognitive processes dissociated from object-level ones

Early evidence for specific brain regions being involved in metacognitive 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 FOK 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. These unique impairments in metamemory performance compared to other amnesic patients suggest that the deficits do not

result from general memory impairments. Furthermore, supporting this separation between object-level and metacognitive processes, Chua and Ahmed (2016) found that transcranial direct current stimulation (tDCS) over dorsolateral PFC improved FOK judgements while leaving task performance unaffected. Kao et al. (2005) presented further evidence that distinguishes brain regions concerned with retrieving a memory from regions monitoring confidence in those memories. Activity in ventromedial PFC varied with participants' predictions of whether they would remember presented 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 in a different task domain. Correlating structural magnetic resonance imaging (MRI) data with task performance, they found that grey-matter volume in a region in right anterior PFC predicted participants' metacognitive sensitivity in a perceptual decision task after controlling for task performance. Meanwhile, studies leveraging the temporal resolution of electroencephalography (EEG) have shown that people's reported confidence in their decisions depends on continued processing after the initial choice, and that this post-decision processing is reflected in neural activity distinct from 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. 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 recalling 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 FOK judgements in patients with damage in the lateral PFC performing a face memory task. 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 the separation of metacognitive judgements based on their temporal focus, there is a debate about whether separate neural substrates support metacognitive judgements across cognitive domains (Vaccaro and Fleming, 2018). One study assessed whether lesions to anterior PFC affected metacognition in both perception and memory (Fleming et al., 2014). While 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 idea is further supported by neuroimaging studies in healthy participants, using resting-state connectivity (Baird, 2013), measurements of variation in white matter microstructure (Baird et al., 2015), voxel-based morphometry (McCurdy et al., 2013) and cortical thickness mapping (Valk et al., 2016). Reconciling these findings with studies arguing for domain-general metacognitive processes (de Gardelle and Mamassian, 2014; Fleck et al., 2006; Schraw, 1996), Morales et al. (2018) proposed the coexistence of both 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. Faivre et al. (2018) found that metacognitive efficiency was correlated across tactile, auditory, visual, and audio-visual 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. Such a *"common currency"* representation of confidence estimate 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

Another pattern that has emerged is a distinction between posterior/medial and anterior/lateral 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). 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 post-decision evidence 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 tracked the impact of post-decision evidence on subjective confidence judgements.

Consistent with these findings, Qiu et al. (2018) suggested that there may be two stages of metacognitive processes: the first stage coincides with object-level processes and generates generic uncertainty information, whereas the second stage transforms this information into a general sense of confidence. This hypothesis 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, supporting that these regions encode the strength of sensory evidence. By contrast, anterior PFC stimulation affected metacognitive sensitivity, indicating that these regions use information about signal strength to generate explicit confidence judgements. Going back to the distinction between implicit and explicit metacognition, 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 use of confidence judgements (Bang et al., 2020; Gherman and Philiastides, 2018).

Neural substrates of metacognitive monitoring?

The neuropsychological and neuroimaging studies reviewed reveal consistent involvement of prefrontal regions in metacognitive processes, with rostrolateral, ventromedial, and dorsomedial regions implicated in many studies (Figure 4). However, no single region has been consistently identified and hypothesised functional sub-divisions are yet to be fully replicated. In other

words, different metacognitive experiences may be supported by distinctive brain regions. For instance, Maril et al. (2005) found increased ACC activity for TOT feelings but not for FOK judgments, suggesting a qualitative difference in these metacognitive states rather than TOT merely being a stronger instantiation of FOK. Regions outside of the core network shown in Figure 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 (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).

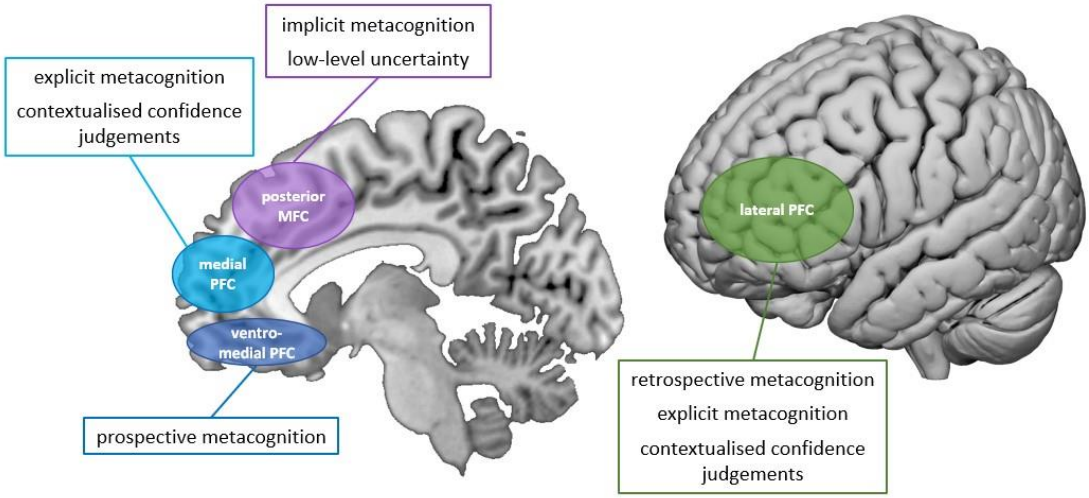

Figure 4. Medial (left) and lateral (right) view of the human brain highlighting hypothesised functional roles in metacognition

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.

This heterogeneity raises questions about our conceptual understanding of metacognition and 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 judgments may depend on multiple cues and be used flexibly depending on the situation, then perhaps finding activity in different areas may be just what we would expect (Schwartz and Díaz, 2014). If so, determining the neural substrates of metacognition requires first disentangling the various processes that underlie a given metacognitive judgement.

Neural substrates of metacognitive control?

The importance of metacognition 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 that someone else's thoughts are inserted into their minds (Ratcliffe and Wilkinson, 2015). Indeed, there is an association between psychiatric symptom dimensions and metacognitive distortions (Rouault et al., 2018b). Such lack of insight can lead to lower medication adherence and poor treatment outcomes (Novick et al., 2015). Yet, despite the established significance of (mal)adaptive metacognitive control, we know much less about its neural substrates, with most focus to date being on metacognitive monitoring processes.

Metacognition guiding adaptive behaviour

The interplay between metacognitive monitoring and metacognitive control processes is important to initiate self-regulatory actions, learn from mistakes, and optimise behaviour. In memory, FOKs prompt sustained attempts at memory retrieval (Costermans et al., 1992), JOLs guide choices about which information should be refreshed and restudied (Kornell and Metcalfe, 2006), and general evaluations of our own mental abilities and their shortcomings play an integral role in deciding when to set external reminders (Gilbert, 2015).

Metacognitive control plays an integral role in allowing us to allocate study time efficiently during learning. Examining metacognitively guided study time allocation, Dunlosky and Herzog (1998) proposed the discrepancy-reduction model. According to this model, people prioritize studying the most difficult items in order to reduce the largest discrepancies from their learning criterion and persevere until an item reaches their internal criterion of being sufficiently well learned. However, following observations of people preferring to study items of easy and medium difficulty and opposing the discrepancy-reduction model, Metcalfe and

Kornell (2005) proposed the Region of Proximal Learning model. This model proposes that people selectively allocate their study time according to their judgement of the rate of learning: After eliminating already mastered material, people prioritize items from easiest to most difficult, allocating study time to an item as long as they perceive themselves to be making progress.

In decision-making, 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. This adaptive behaviour is also seen in animal studies. For instance, Kepecs et al. (2008) observed that rats adapted their waiting time in accordance with how confident they were in obtaining a reward.

Metacognitive processes not only improve immediate decision-making but also shape behavioural strategies for future decisions, for example by adopting a more cautious approach following errors (Rabbitt and Rodgers, 1977) and low confidence decisions (Van den Berg et al., 2016). 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 et al., 2019) or unacquainted tasks (Rouault et al., 2018a).

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 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 (so far somewhat neglected) aspect of metacognition, as well as a potentially critical factor producing variability in activation foci across studies with different methodologies.

Metacognition in a social world

As well as guiding individuals' adaptive behaviour, metacognition plays an important functional role in allowing us to report our experiences and justify our own behaviour to others (Frith, 2012). Being able to reflect on our own cognitive processes and mental states may thus enhance social interactions. Frith and Frith (2012) pointed out that one's mental representations of the world are representations only, and that metacognition enables us to understand that others may have different representations (with the possibility that our representations may be inaccurate). Such insights allow us to discuss and compare our reflections and to build a shared perspective.

Sharing metacognitive evaluations with others can facilitate adaptive group behaviour (Shea et al., 2014). 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). For instance, 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). Bahrami et al. (2010) showed

that dyadic decision-making performance can be described by a weighted-confidence sharing model, whereby the individual responses are weighted by each agent's confidence to arrive at a joint decision. Indeed, 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; cf. Bang et al., 2017 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).

However, it remains unknown whether, or to what extent, the social sharing of metacognitive representations can improve individual-level task performance and metacognition. The social context can also influence an individual's self-directed metacognitive appraisals. For instance, Wittmann et al. (2016) found that interacting with an objectively better partner led to an overestimation of one's own abilities in a cooperative context but an underestimation of one's own abilities in a competitive context. Notably, the opposite pattern was found for an objectively worse partner. Studying TOT states, Rousseau and Kashur (2021) found that people experience more TOTs when trying to remember something in small groups as opposed to by themselves, suggesting people may (unconsciously) infer that successful retrieval may be more likely as a group than alone.

How is metacognition represented in the brain?

As indicated above, PFC seems to play a critical role in supporting metacognitive processes but, within this broad region of the cerebral cortex, with different studies reporting 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 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.

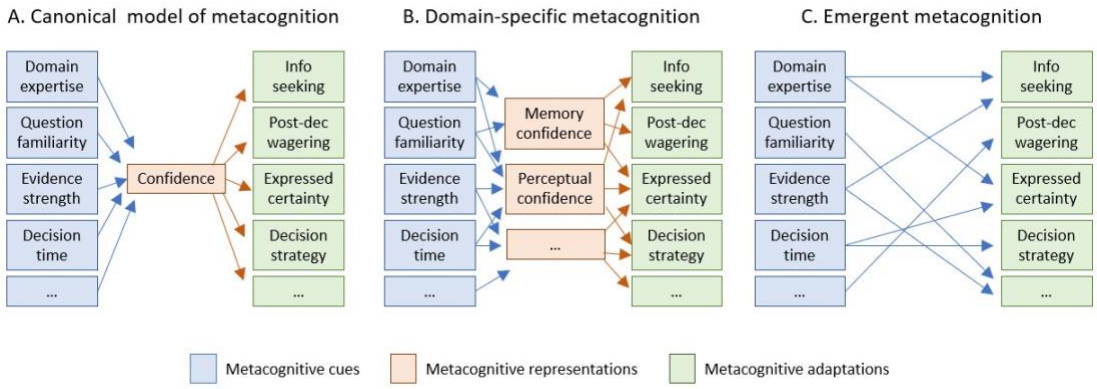

Figure 5. Models of metacognition

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 (Figure 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 accompanied by some subjective experience of metacognition, without the need for a common representation or

general mediation (Figure 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, how it relates to subjective metacognitive experiences, and where it is instantiated in the brain.

Relationship between metacognition and Theory of Mind?

Having a Theory of Mind (ToM, also called mentalizing) refers to the ability to infer mental states (e.g., beliefs, desires, intentions) and predict behaviours of others (Premack and Woodruff, 1978). This is classically tested using false belief tasks (Wimmer and Perner, 1983), which requires understanding that others' beliefs may be incorrect and different to one's own. The notion of understanding mental states and predicting or influencing behaviours of others exhibits striking similarities to monitoring and controlling one's own cognitions (i.e., metacognition). Indeed, Frith (2012) noted that, *"understanding of behaviours in terms of beliefs and desires is an example of explicit metacognition, whether it is applied to the self or to others"*. If so, then a common computational basis and shared supporting neural processes may be expected.

Indeed, individual differences studies suggest a positive relation between ToM and metacognitive abilities (van der Plas et al., 2021). Moreover, in their meta-analysis of neuroimaging studies, Fleming et al. (2018) report, alongside unique activation associated with each function, an overlap in neural activity patterns related to metacognition and ToM in the ventromedial PFC and dorsomedial PFC. Based on these findings, Fleming et al. (2018) suggest that brain areas uniquely linked to metacognition may process lower-level epistemic feelings of uncertainty, while activation shared with ToM may be related to more conscious higher-order representations of mental states. Hence, explicit, as opposed to implicit, metacognition may rely

on shared neural mechanisms with ToM. Indeed, Jenkins et al. (2008) used repetition suppression, a process by which neural activity is typically reduced following repeated presentation of conceptually related constructs, to show attenuated activity in the ventromedial PFC for introspection about the self and reflections about similar others following a preceding self-reflection task. However, this repetition suppression was not observed for reflections about dissimilar others. Thus, the relation between metacognition and ToM might depend on the target of mentalizing with metacognitive processes concerned with introspection only being recruited for inferring mental states of sufficiently similar others (Mitchell et al., 2005, 2006).

Further evidence for a close relationship between metacognition and ToM comes from the study of various psychopathologies. Autistic traits are associated with deficiencies in inferring mental states of others (Baron-Cohen et al., 2001) and have also been linked to impaired metacognitive abilities (Grainger et al., 2014). Similarly, obsessive-compulsive disorder (OCD), associated with distortions in metacognitive functions (Janeck et al., 2003), has also been implicated with weak performance in ToM tasks (Yazici and Yazici, 2019). Further, Thibaudeau et al. (2017) showed that a therapeutic program targeting metacognitive functions lead to improvements in ToM abilities for patients diagnosed with schizophrenia, a condition associated with deficiencies in internal monitoring (Frith and Done, 1988).

It remains to be determined how the capacity to reflect on our own thoughts relates to our capacity to infer the mental states of others, and whether these abilities depend on common neurocognitive systems (Carruthers, 2009). For example, ToM and metacognition could be based on a single cognitive mechanism with two kinds of access. Metacognition may be prior whereby the mental states of others are inferred by simulation based on one's own mind. Conversely, mentalizing may be prior whereby metacognition is a self-directed inference based

on ToM. Further explorations of these links may contribute to a better understanding of the processes underlying metacognition.

What is the nature of metacognitive abnormalities in psychological disorders?

A range of psychiatric conditions are characterized by distortions in metacognition suggesting that deeper insights into the processes underlying metacognition may be valuable in better understanding psychopathology and developing treatment interventions (Rouault et al., 2018b). For instance, depression has been linked to overall decreased levels of confidence (Fu et al., 2012), patients with schizophrenia display diminished metacognitive sensitivity (Eifler et al., 2015), and clinical addiction and OCD have been associated with decreased metacognitive efficiency (Hauser et al., 2017; Moeller et al., 2016). Metacognitive abnormalities could be a result or mere by-product 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.

Some causal evidence comes from longitudinal studies looking at symptom severity in relation to the degree to which metacognition is compromised. 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; Gawęda et al., 2018). Within clinical populations, a correlation between levels of metacognitive abnormalities and symptom severity has been reported. Hancock (1996), found confidence in correct choices to negatively correlate with depression severity and Nicolò et al. (2012) report 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 (van den Hout and Kindt, 2003) and, conversely, undermining confidence can foster compulsive tendencies (Cuttler et al., 2013).

Building on this body of research, current findings may be translated into interventions tailored to individual deficits. Metacognition has been a target of therapeutic interventions for various mental disorders (Moritz and Woodward, 2007; Wells, 2011). However, considering high comorbidity rates as well as considerable symptom variability within individual diagnoses, interventions may usefully focus on metacognitive abnormalities as related to symptom dimensions, adopting a transdiagnostic approach instead of categorical disorder classifications (transdiagnostic psychiatry – see Fusar-Poli et al., 2019). Moreover, the neurological mechanisms underlying symptom improvement following metacognitive therapy remain poorly understood.

Summary and Conclusions

This chapter explores the neural underpinnings of metacognition – the monitoring and controlling of one's own thoughts. Reviewing studies employing diverse neuroimaging techniques, we have discussed the functional localization as well as our theoretical understanding of metacognition. After considering the nature of metacognition and its measurement, we have discussed whether there are meaningful explicit representations of metacognitive signals, using 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. We have considered the functional value of explicit metacognitive signals, in particular the social sharing of such estimates. Finally, we have

reviewed debates that cognitive neuroscience may help address, focusing on the relationship between metacognition and theory of mind, and the role of metacognition in the treatment of psychiatric disorders.

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. Anterior/lateral regions of PFC are consistently implicated in these processes. However, 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 neural 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.

REFERENCES

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), 3–8.

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

Bang, D., Aitchison, L., Moran, R., Castanon, S. H., Rafiee, B., Mahmoodi, A., ...Summerfield, C. (2017). Confidence matching in group decision-making. *Nature Human Behaviour*, *1*. doi:10.1038/s41562-017-0117

Bang, D., Ershadmanesh, S., Nili, H., & Fleming, S. M. (2020). Private-public mappings in human prefrontal cortex. *Elife*, *9*. doi:10.7554/eLife.56477

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

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The “reading the mind in the eyes” test revised version: A study with normal adults, and adults with

Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *42*(2), 241–251.

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

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*. doi: 10.1038/s41598-019-40681-9

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.

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), 729–748.

Carruthers, P. (2009). How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences*, *32*(2), 121–138.

Chua, E. F., & Ahmed, R. (2016). Electrical stimulation of the dorsolateral prefrontal cortex improves memory monitoring. *Neuropsychologia*, *85*, 74–79.

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), 142–150.

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.

de Gardelle, V., Le Corre, F., & Mamassian, P. (2016). Confidence as a common currency between vision and audition. *Plos One*, *11*(1). doi: 10.1371/journal.pone.0147901

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

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.

Dunlosky, J., & Hertzog, C. (1998). Training programs to improve learning in later adulthood: Helping older adults educate themselves. In D. J. Hacker, J. Dunlosky & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 249–275). New York: Routledge.

Eifler, S., Rausch, F., Schirmbeck, F., Veckenstedt, R., Mier, D., Esslinger, C., ...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., ...Zink, M. (2015). Investigation of metamemory functioning in the at-risk mental state for psychosis. *Psychological Medicine*, *45*(15), 3329–3340.

Faivre, N., Filevich, E., Solovey, G., Ku¨hn, 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.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist, 34*(10), 906–911.

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), 91–114.

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

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience, 8*. doi: 10.3389/fnhum.2014.00443

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,* 1541–1543.

Friedman, N. P., & Robbins, T. W. (2022). The role of prefrontal cortex in cognitive control and executive function. *Neuropsychopharmacology*, *47*(1), 72–89.

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

Frith, C. D., & Done, D. J. (1988). Towards a neuropsychology of schizophrenia. *British Journal of Psychiatry*, *153*(4), 437–443.

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

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.

Fusar-Poli, P., Solmi, M., Brondino, N., Davies, C., Chae, C., Politi, P., ...McGuire, P. (2019). Transdiagnostic psychiatry: A systematic review. *World Psychiatry*, *18*(2), 192–207.

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.

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

Gherman, S., & Philiastides, M. G. (2018). Human VMPFC encodes early signatures of confidence in perceptual decisions. *Elife*, *7*. doi: 10.7554/eLife.38293

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.

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.

Grainger, C., Williams, D. M., & Lind, S. E. (2014). Metacognition, metamemory, and mindreading in high-functioning adults with autism spectrum disorder. *Journal of Abnormal Psychology*, *123*(3), 650–659.

Green, J., Goldstein, F. C., Sirockman, B. E., & Green, R. C. (1993). Variable awareness of deficits in Alzheimer's disease. *Neuropsychiatry, Neuropsychology, & Behavioral Neurology*, *6*(3), 159–165.

Guggenmos, M., Wilbertz, G., Hebart, M. N., & Sterzer, P. (2016). Mesolimbic confidence signals guide perceptual learning in the absence of external feedback. *Elife*, *5*. doi: 10.7554/eLife.13388.001

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

Hauser, T. U., Allen, M., Rees, G., & Dolan, R. J., Bullmore, E. T., Goodyer, I., …Pantaleone, S. (2017). Metacognitive impairments extend perceptual decision-making weaknesses in compulsivity. *Scientific Reports*, *7*(1). doi: 10.1038/s41598-017-06116-z

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.

Janeck, A. S., Calamari, J. E., Riemann, B. C., & Heffelfinger, S. K. (2003). Too much thinking about thinking? Metacognitive differences in obsessive-compulsive disorder. *Journal of Anxiety Disorders*, *17*(2), 181–195.

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.

Jenkins, A. C., Macrae, C. N., & Mitchell, J. P. (2008). Repetition suppression of ventromedial prefrontal activity during judgments of self and others. *Proceedings of the National Academy of Sciences*, *105*(11), 4507–4512.

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

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

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

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.

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), 34–53.

Koriat, A. (2007). Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *The Cambridge handbook of consciousness* (pp. 289-325). Cambridge, UK: Cambridge University Press.

Kornell, N., & Metcalfe, J. (2006). Study Efficacy of Proximal Learning Framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(3), 609–622.

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.

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). doi: 10.1038/s42003-020-1049-3

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

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

Maril, A., Simons, J. S., Weaver, J. J., & Schacter, D. L. (2005). Graded recall success: An event-related fMRI comparison of tip of the tongue and feeling of knowing. *Neuroimage, 24*, 1130–1138.

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.

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.

Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of Memory and Language, 52*(4), 463–477.

Metcalfe, J., & Schwartz, B. L. (2016). The ghost in the machine: Self-reflective consciousness and the neuroscience of metacognition. In J. Dunlosky & S. Tauber (Eds), *Oxford Handbook of Metamemory* (pp. 407–424). Oxford, UK: Oxford University Press.

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

Mitchell, J. P., Banaji, M. R., & Macrae, C. N. (2005). The link between social cognition and self-referential thought in the medial prefrontal cortex. *Journal of Cognitive Neuroscience, 17*(8), 1306–1315.

Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron*, *50*(4), 655–663.

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*, 188–193.

Moeller, S. J., Fleming, S. M., Gan, G., Zilverstand, A., Malaker, P., d'Oleire Uquillas, F., ...Goldstein, R. Z. (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., & Woodward, T. S. (2007). Metacognitive training in schizophrenia: From basic research to knowledge translation and intervention. *Current Opinion in Psychiatry, 20*(6), 619–625.

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.* doi: 10.7554/eLife.11946

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95*(1), 109–133.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *Psychology of learning and motivation* (Vol. 26, pp. 125-173). New York: Academic.

Nicolo`, G., Dimaggio, G., Popolo, R., Carcione, A., Procacci, M., Hamm, J., ...Lysaker, P. H. (2012). Associations of metacognition with symptoms, insight, and neurocognition in

clinically stable outpatients with schizophrenia. *Journal of Nervous and Mental Disease*, *200*(7), 644–647.

Novick, D., Montgomery, W., Treuer, T., Aguado, J., Kraemer, S., & Haro, J. M. (2015). Relationship of insight with medication adherence and the impact on outcomes in patients with schizophrenia and bipolar disorder: Results from a 1-year European outpatient observational study. *BMC Psychiatry*, *15*(1). doi: 10.1186/s12888-015-0560-4

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.

Pescetelli, N., Hauperich, A.-K., & Yeung, N. (2021). Confidence, advice seeking and changes of mind in decision making. *Cognition*, *215*. doi: 10.1016/j.cognition.2021.104810

Pescetelli, N., Rees, G., & Bahrami, B. (2016). The perceptual and social components of metacognition. *Journal of Experimental Psychology: General*, *145*(8), 949–965.

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

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515–526.

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). doi: 10.1371/journal.pbio.2004037

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.

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

Reder, L. M., & Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*, 435–451.

Rhodes, M. G. (2015). Judgments of learning: Methods, data, and theory. In J. Dunlosky & S. (Uma) K. Tauber (Eds.), *The Oxford handbook of metamemory* (pp. 65-80). Oxford, UK: Oxford University Press.

Rouault, M., McWilliams, A., Allen, M. G., & Fleming, S. M. (2018a). Human metacognition across domains: Insights from individual differences and neuroimaging. *Personality Neuroscience, 1*. doi: 10.1017/pen.2018.16

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

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*. doi: 10.3389/fpsyg.2021.704433

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.

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. *Journal of Experimental Education*, *65*(2), 135–146.

Schwartz, B. L., & Díaz, F. (2014). Quantifying human metacognition for the neurosciences. In S. M. Fleming & C. D. Frith (Eds.), *The cognitive neuroscience of metacognition* (pp. 9-23). Berlin: Springer-Verlag.

Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Suprapersonal 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). doi: 10.1093/nc/niw005

Shekhar, M., & Rahnev, D. (2018). Distinguishing the roles of dorsolateral and anterior PFC in visual metacognition. *Journal of Neuroscience*, *38*(22), 5078–5087.

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), 452–460.

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.

Son, L. K., & Metcalfe, J. (2005). Judgments of learning: Evidence for a two-stage process. *Memory & Cognition*, *33*(6), 1116–1129.

Terrace, H. S., & Son, L. K. (2009). Comparative metacognition. *Current Opinion in Neurobiology*, *19*(1), 67–74.

Thibaudeau, E., Cellard, C., Reeder, C., Wykes, T., Ivers, H., Maziade, M., ...Achim, A. M. (2017). Improving theory of mind in schizophrenia by targeting cognition and metacognition with computerized cognitive remediation: A multiple case study. *Schizophrenia Research and Treatment, 2017*. doi: 10.1155/2017/7203871

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*. doi: 10.1177/2398212818810591

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 Plas, E., Mason, D., Livingston, L. A., Craigie, J., Happ´e, F., & Fleming, S. (2021). Computations of confidence are modulated by mentalizing ability. *bioRxiv*. doi: 10.31234/osf.io/c4pzj

Wells, A. (2011). *Metacognitive therapy for anxiety and depression*. Guilford press.

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.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103–128.

Yazici, K. U., & Yazici, I. P. (2019). Decreased theory of mind skills, increased emotion dysregulation and insight levels in adolescents diagnosed with obsessive compulsive disorder. *Nordic Journal of Psychiatry*, *73*(7), 462–469.

Yeung, N. (2014). Conflict monitoring and cognitive control. In K. N. Ochsner & S. M. Kosslyn (Eds,), *The Oxford handbook of cognitive neuroscience, Vol. 2: The cutting edges*. (pp. 275-299). Oxford, UK: Oxford University Press.

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*, 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.

Zimmerman, B. J. (1986). Becoming a self-regulated learner: Which are the key subprocesses? *Contemporary Educational Psychology*, *11*(4), 307–313.