METACOGNITION AND CONFIDENCE: A REVIEW AND SYNTHESIS

Stephen M. Fleming
University College London
stephen.fleming@ucl.ac.uk

Forthcoming in *Annual Review of Psychology*
https://annualreviews.org

July 6, 2023

ABSTRACT

Determining the neural basis of confidence and uncertainty holds promise for understanding foundational aspects of human metacognition. While a neuroscience of confidence has focused on the mechanisms underpinning subpersonal phenomena such as representations of uncertainty in the visual or motor system, metacognition research has been concerned with personal-level beliefs and knowledge about self-performance. I provide a roadmap for bridging this divide by focusing on a particular class of confidence computation: propositional confidence in one's own (hypothetical) decisions or actions. Propositional confidence is informed by the observer's models of the world and their cognitive system, which may be more or less accurate – thus explaining why metacognitive judgments are both inferential and sometimes diverge from task performance. Disparate findings on the neural basis of uncertainty and performance monitoring are integrated into a common framework, and a new understanding of the locus of action of metacognitive interventions developed.

1 INTRODUCTION

Imagine you are revising for an upcoming exam in psychology. At various points leading up to the big day, you wonder whether you know the material well enough, or not. Such an assessment might prompt further study, until those uncertainties are diminished, and you feel more confident in being able to answer anything that is thrown at you. Before going into the exam hall, you nervously compare your chances of success with your friends. Later, after the exam is over, you think back over your answers, questioning whether it went well, or whether it could have gone better. These forms of self-evaluation are instances of metacognition – the capacity to reflect on, evaluate and control mental function in a variety of useful ways.

These are examples of metacognition about memory, or metamemory for short. But metacognition operates over a range of domains. Consider a visit to the opticians for a new pair of glasses. In a typical eye exam, you will be asked whether you are seeing the world more or less clearly through different lenses. This a metacognitive judgment about your perceptions – the world is not blurry, but a limit on your visual acuity makes it seem so.

It is hopefully clear from these two examples that the accuracy of metacognition – whether or not our self-evaluative judgments match up with the reality of cognitive or physical performance – is central to adaptive behaviour. If I think that my knowledge about a topic is secure when it is in fact shaky, I might put down the books and go out with friends, only to be in with a nasty shock on exam day. Similarly, if we are unable to realise when our vision (or hearing, or memory) is failing, we will be unable to take steps to correct for physical or cognitive limitations. As such, metacognitive dysfunction has been highlighted as a key source of maladaptive behaviour in educational, clinical and societal contexts (Flavell, 1979, Hoven et al., 2019, Rollwage et al., 2018).

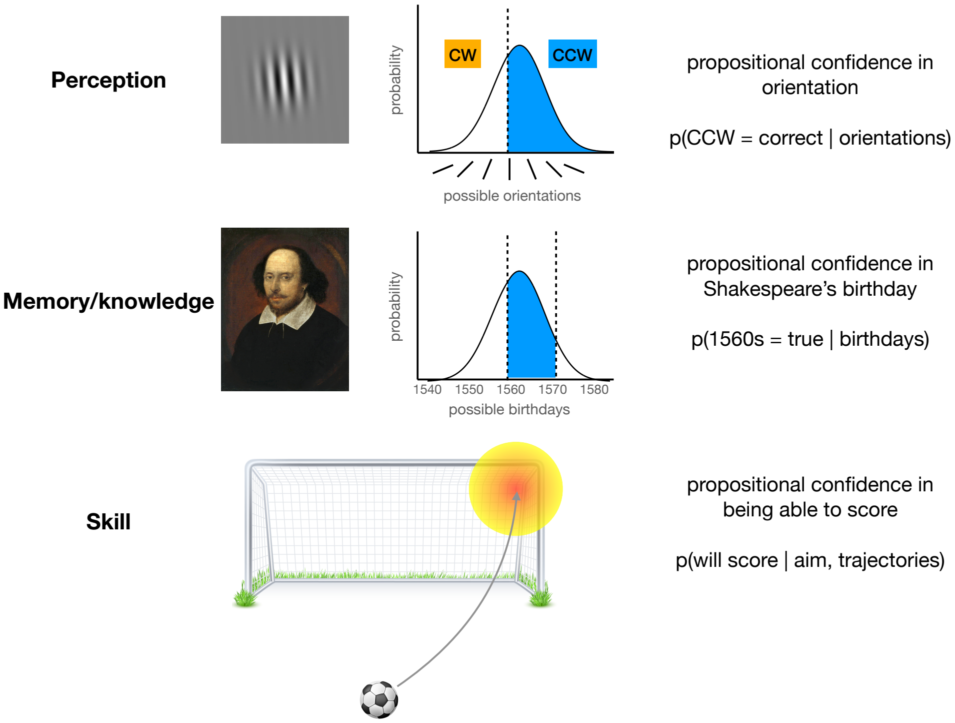
Perception
CW
CCW
probability
possible orientations
propositional confidence in
orientation
p(CCW = correct | orientations)
Memory/knowledge
probability
1540
1550
1560
1570
1580
possible birthdays
propositional confidence in
Shakespeare's birthday
p(1560s = true | birthdays)
Skill
propositional confidence in
being able to score
p(will score | aim, trajectories)

Figure 1: Metacognitive judgments can be formalised as estimates of propositional confidence across a range of domains and timescales.

Effectively estimating our uncertainty or confidence in a range of cognitive processes, and whether or not such confidence judgments track objective performance (known as metacognitive sensitivity), is therefore central to effective metacognition (Nelson & Narens, 1990). Miscalibrated confidence in success can lead to failure, even when our natural aptitude is more than adequate. Recently there has been surge of interest in the neuroscience of uncertainty and confidence, leading to a marriage of computational work in cognitive science with human neuroimaging studies and animal models of metacognitive judgments (Pouget et al., 2016, Meyniel et al., 2015). Partly because these fields were steeped in the methods of psychophysics, and partly because of the cross-species tractability of perceptual paradigms, the late 2000s saw the emergence of the field of perceptual (largely visual) metacognition, with a strong focus on the neural and computational underpinnings of confidence judgments (Rahnev, 2021).

However, the rapid rise of this research program brings with it a set of pressing conceptual challenges. The neuroscience of confidence has tended to focus on the mechanisms underpinning subpersonal phenomena such as the representation of uncertainty in the visual or motor system, often in tightly controlled laboratory tasks. Conversely, metacognition researchers are interested in personal level beliefs and knowledge in real-world settings: why do I think that I performed poorly on the exam? How do I recognize when I might have made a poor decision? Why is a patient with Alzheimer's disease unaware of their memory failures? How do children form beliefs about what they know and do not know?

In this article I aim to provide a roadmap for bridging this divide. Metacognition and confidence researchers are natural allies, but have often been uneasy bedfellows, with the latter thinking that the former are overcomplicating things, and the former thinking the latter are riding roughshod over the richness of metacognition by reducing it down to its computational primitives. I suggest that one solution to understanding the role of confidence in real-world metacognition is to focus on a particular class of confidence computation – propositional confidence. Propositional confidence is confidence in one's own (hypothetical) decisions or actions – which include covert propositions ("I think I will remember this word"; **Figure 1**). The most important idea, building on Pouget et al. (2016), is that propositional confidence can be distinguished from a myriad of other confidences or uncertainties that are inherent to perception, cognition and action – although the latter often inform the former (Meyniel et al., 2015). Propositional confidence is also affected by the observer's models of the world and their cognitive system, which may be more or less accurate – thus explaining why metacognitive judgments are both inferential and sometimes diverge from task performance.

2 SCOPE AND DEFINITIONS

The terms metacognition and confidence can take on different meanings in different research fields, and so it is useful to spend some time providing explicit definitions.

By "metacognition", I refer to the class of mechanisms that allow us to form beliefs about other mental operations. Such beliefs (the monitoring aspect of metacognition) can then be harnessed for self-regulation (metacognitive control) and/or for communicating metacognitive assessments to others. Metacognition is a part of the wider set of human executive functions, although conceptually and empirically distinct from fluid intelligence: it is possible (and indeed common) to evaluate the operation of classical executive functions, for instance reflecting on whether a solution to a logical puzzle was in fact appropriate (Ackerman & Thompson, 2017). The accuracy of such reflective judgments shares variance with other forms of metacognitive sensitivity, rather than variance in IQ (Mazancieux et al., 2020, Rouault et al., 2018a). Finally, metacognition also intersects with the literature on cognitive control, although again with only partial overlap. Cognitive control typically refers to the set of functions that encode and maintain a representation of the current (first-order) task. For instance, in Miller & Cohen (2001)'s classic model of cognitive control, information in prefrontal cortex provides contextual signals to bias or route sensory information to establish the right mapping between inputs, internal states and outputs. All of this machinery can be considered as being part of the same (context-sensitive) first-order system. We can then apply metacognitive mechanisms to monitor task performance and subsequently increase our reliance on cognitive control (Norman & Shallice, 1986). The literature on error correction and performance monitoring has often been lumped together with the literature on cognitive control, but here would also fall under the rubric of metacognition research.

By "confidence", I mean the degree of belief one has about the likely success of a variety of mental operations. Thus confidence here refers to propositional confidence – a feeling of surety about your abilities, judgments or ideas. Confidence also has a more general meaning as a synonym with probability: as in, ascribing a high probability (high confidence) that the sun will rise tomorrow. Such probabilities apply to external quantities, independently of an observer. To add to the confusion, it is also possible that the brain itself uses probabilistic computation in a range of processes, including the formation of feelings of confidence! To try to avoid confusion here I will follow Pouget et al. (2016) and reserve the term confidence to refer to propositional confidence in a (mental or physical) action; and use the term "certainty" (or its converse, uncertainty) to refer to degree of belief in other quantities.

I aim to bridge between work on subpersonal representations of uncertainty, personal-level feelings of confidence, and the operation of metacognition more broadly. This necessarily means being selective in the empirical literature that is most helpful in illuminating those relationships. As such, there are a number of topics that fall outside the scope of the review, given limited space. These are: the development of metacognition; comparative research on animal metacognition; links between metacognition, mental health and ageing (although see sidebar on Group and Individual Differences); and interpersonal and intrapersonal functions of metacognition.

The outline of the article is as follows. In Section 3 I provide a brief overview of core findings in metacognitive neuroscience that motivate the current synthesis. Section 4 then deconstructs the different components of a personal-level metacognitive judgment, and reviews the evidence for distinct components, with a particular focus on neuroscience. An important concept here will be the notion of a "reference frame". We can talk of uncertainty about things in the world – such as sensory uncertainty about the orientation of a line, or the frequency of a sound. This is uncertainty in a "world-centred" reference frame. But as we have seen, we can also talk of confidence in our own propositions or actions – this is now uncertainty in a "self-centred" reference frame. I then turn to how such signals are read out or broadcast in a format that is useful for guiding behaviour and communication to others, before evaluating the role that model-based computation plays in providing contextual knowledge for metacognition.

The remainder of the paper asks how current controversies in metacognition research can be re-evaluated in light of this framework – in particular, the origin of biases and suboptimalities in metacognition, how to arbitrate between computational models of confidence, and whether or not we should consider metacognition as a domain-general resource. I close by highlighting some future directions that are motivated by this framework – in particular, searching for common computational principles across different task domains; how we might extend models of local confidence formation to understand the formation of metacognitive knowledge over longer timescales; and identifying the best routes for interventions on metacognition.

3 PARADIGMS AND FINDINGS IN METACOGNITIVE NEUROSCIENCE

A range of behavioural paradigms investigating different types of metacognitive judgment have been devised, often originating in work on metamemory – ranging from prospective "judgments of learning" to retrospective confidence estimates in recall (Metcalfe & Shimamura, 1994). All, however, have in common that subjects are being asked to evaluate their (future or past) performance on another task. As we will see, such evaluations are naturally cast as judgments of propositional confidence in the success of other mental operations. In humans, these judgments are usually explicit and instructed – subjects are provided with a button or scale on which to indicate their confidence, or are asked, in the confidence forced-choice paradigm, to pick from a pair of decisions the one they feel most confident about (Mamassian & de Gardelle, 2022). In animal metacognition research, confidence estimates are elicited using a variety of learnt second-order contingencies such as opting out of a decision, waiting for a reward that is contingent on first-order task performance, and so on (Kepecs & Mainen, 2012). These so-called "implicit" measures of metacognition have recently found their way into innovative studies of infant metacognition, where explicit confidence elicitation is less straightforward (Goupil & Kouider, 2016).

When we have data on a series of metacognitive judgments over time, we can examine the statistical association between behavioural performance and metacognition. Intuitively, if you are confident when you are right, and less confident when you are wrong, then you can be ascribed a high degree of metacognitive *sensitivity* (Fleming & Lau, 2014). Another relevant summary statistic for investigations of metacognition is metacognitive *bias* (also known as calibration, or overconfidence) – the extent to which subjects tend to report higher or lower confidence, relative to long-run performance. One challenge is to ensure measures of metacognitive sensitivity are unconfounded by other factors, including task performance, metacognitive biases, and response times (see Sidebar).

Measurement of metacognition

Measures of metacognition in experimental tasks seek to estimate the statistical relationship between confidence judgments and objective performance, known as metacognitive sensitivity. A central challenge in this endeavour is to ensure metrics of metacognitive sensitivity are unconfounded by other influences. For instance, simple correlations between accuracy and confidence depend not only on metacognitive sensitivity but are also affected by performance and metacognitive bias (average confidence level; Fleming & Lau, 2014). The meta-*d'* model offers a performance-controlled metric of metacognitive sensitivity, by estimating the level of first-order performance (*d'*) that would have given rise to the observed confidence data under a signal detection theoretic model (Maniscalco & Lau, 2012). The ratio meta-*d'/d'* thus provides a performance-controlled metric of metacognitive capacity (often referred to as metacognitive *efficiency*). However, the assumption that meta-*d'/d'* is fully independent of confidence and performance has been challenged (Guggenmos, 2021, Xue et al., 2021). Alternative model-free approaches assess the mutual information between performance accuracy and confidence reports (Dayan, 2022) or quantify the change in psychometric function slope as a function of confidence (De Martino et al., 2013, de Gardelle & Mamassian, 2014).

With these metrics in place, two lines of work have emerged in metacognitive psychology and neuroscience over the past few decades. The first has sought to catalogue both individual differences and interventions – either experimentally-controlled, or naturally occurring in the form of brain damage or disorder – that affect metacognition without affecting first-order task performance. A second line of work has focused on the psychological, computational and neural basis of confidence formation across a number of different task domains, in both humans and animal models. Classical work in the cognitive psychology of metamemory has identified a range of cues that may affect confidence judgments, but are unrelated to first-order performance. For instance, when attempting to recall a difficult-to-retrieve item, the extent to which we can recall information related to the target (cue accessibility) predicts how confident we are in being able to recognize the target (Koriat, 1993). A number of these influences on metamemory judgments have been studied in depth – including target accessibility, fluency at encoding and retrieval, and response time – leading to the broad proposal (that we will return to below) that metacognitive judgments are inferential in nature, and draw on a range of helpful and unhelpful cues to performance (Metcalfe & Shimamura, 1994, Nelson & Narens, 1990). Within the field of metaperception research, studies have documented dissociations between confidence and accuracy as a function of attention (e.g., Wilimzig et al., 2008), variability in perceptual evidence (e.g., Zylberberg et al., 2014, Spence et al., 2016), asymmetries in the processing of supporting and disconfirming evidence (e.g., Zylberberg et al., 2012, Miyoshi & Lau, 2020), and response times (e.g., Kiani et al., 2014). These research programs on individual differences and confidence formation naturally reinforce one another, as new discoveries about the formation of confidence can shed light on the origins of individual and group differences, and identifying individual and group differences in metacognitive efficiency provides hints about where to look for sources noise or suboptimality in confidence formation.

Pioneering neuropsychological investigations of patients with frontal lobe damage have identified a key role for the human prefrontal cortex in supporting metacognitive capacity, often on memory tasks (see Pannu & Kaszniak, 2005, Fleming & Dolan, 2012, for reviews). The importance of prefrontal cortical function in metacognition has been supported by recent studies in both humans and animals. Changes in confidence formation and metacognition (but not first-order task performance) are observed following temporary disruption or lesions to rostrolateral prefrontal cortex (Brodmann areas 46 and 10) in humans and monkeys (Fleming et al., 2014, Shekhar & Rahnev, 2018, Miyamoto et al., 2017, 2018, Kwok et al., 2019), and confidence-related behaviour is impaired following inactivation of orbitofrontal cortex in rodents (Lak et al., 2014). Individual differences in perceptual metacognitive sensitivity have been similarly linked to variation in the structure and function of human anterior prefrontal cortex (Fleming et al., 2010, McCurdy et al., 2013, Baird et al., 2013, Allen et al., 2017). This picture of a unitary prefrontal correlate of metacognition has been nuanced with observations in humans that distinct brain systems may predict metacognitive sensitivity in perception and memory tasks (Baird et al., 2013, McCurdy et al., 2013, Fleming et al., 2014, Ye et al., 2018), and that connectivity between prefrontal cortex and other brain areas is important for metacognitive capacity (Baird et al., 2013, 2015, De Martino et al., 2013, Zheng et al., 2021).

Finally, a number of studies in both human and animal models have sought to relate variation in subjective confidence reports, or confidence-related behaviours, to changes in neural activity measured either with single-unit recordings, or mass univariate analyses of neuroimaging data. Many of these studies will be discussed in more detail in subsequent sections. For now, it is sufficient to say that field has catalogued a wide variety of confidence-related neural signals (Walker et al., 2022), with the functional anatomy of metacognition becoming both richer and more complex. Imposing order on these findings is one of the goals of this review: how can we square the often striking dissociations between performance and metacognition observed in lesion studies, with the multiplicity of neural representations of uncertainty and confidence? In the remainder of this article I develop the computational components of a metacognitive judgment, beginning with a theoretical perspective, and then turning to consider the behavioural and neuroscience evidence for each.

4 COMPONENTS OF A METACOGNITIVE JUDGMENT

4.1 Representing uncertainty

Metacognitive assessments refer to one's degree of certainty or uncertainty about a particular mental operation. It is thus natural that the uncertainty inherent to neural representations of sensory features should be highly relevant to metacognition. When a doctor views an X-ray, the incoming visual information may be consistent with a number of different interpretations (both of simple features such as lines and edges, and of more global properties such as the presence or absence of a tumour). It is increasingly recognized that uncertainty is inherent to all stages of neural computation, and that optimal behaviour requires sensitivity to such uncertainty. For instance, when combining information from two different sensory modalities, the normative (Bayesian) solution is to weight the two sources inversely according to their respective uncertainties.

Within perceptual systems, different competing theoretical schemas have been proposed for how the brain represents uncertainty about particular quantities. Consider a judgment of the orientation of a low-contrast grating (**Figure 2**). The sensory data underdetermines the true orientation, leading to uncertainty in the internal representation of orientation z (note this uncertainty is subjective uncertainty in the representation, rather than noise in the stimulus, although the latter may affect the former). We can denote such uncertainty as a (posterior) probability distribution around the most probable orientation. For instance, under a probabilistic population coding model, neurons encode parameters of probability distributions, with different neurons tuned to different stimulus features (such as its orientation or colour), such that a population of such neurons represents a probability distribution over features, given a sensory measurement (Ma et al., 2006). Alternative schemes include sampling-based accounts (where samples from a distribution are accumulated over time in the form of spikes) and summary-statistic accounts in which neuromodulators or other aspects of brain activity carry uncertainty-related information (Fiser et al., 2010, Yu & Dayan, 2005).

For our current purposes, it is sufficient to note that a number of theoretical accounts propose that neural representations come along with an implicit representation of the certainty with which that representation is held. Such distributional uncertainty is thought to be encoded at a number of different levels from perception to cognition and action. As a concrete example, a population of neurons in V1 might (implicitly) carry information about the uncertainty of the orientation of a low-contrast bar; a population of neurons in auditory cortex may carry information about the uncertainty of the frequency of a tone in noise, and so on. These examples hopefully make clear that the brain can and likely does track uncertainty in a whole host of quantities. Bayesian theories of brain function additionally propose that such uncertainties allow the appropriate weighting of messages passed up and down a cognitive hierarchy. Following Meyniel et al. (2015) I refer to these uncertainty signals as implicit or distributional uncertainty, but such estimates may also be transformed into scalar summary signals (for instance, a scalar signal of sensory uncertainty signaled by the level of a particular neuromodulator).

A wide range of studies indicate that subjects take into account uncertainty in their behaviour, including in experiments on perception, learning, memory and motor control (Kersten et al., 2004). Some of the most robust evidence for the representation and use of uncertainty comes from the literature on cue-combination in multisensory integration. If subjects are asked to combine information across two sensory modalities, the weights they put on the two sources of information is inversely proportional to their uncertainty, and approaches the predictions of an ideal Bayesian observer (e.g., Ernst & Banks, 2002). Similarly, in the motor domain, subjects are sensitive to uncertainty in movement production (for instance, the dispersion of rapid pointing movements), and use this information to alter their movement strategies to avoid risky actions (Trommershauser et al., 2008).

Such studies, however, do not tell us whether uncertainty is used to inform metacognition. A number of studies have presented evidence that confidence judgments are sensitive to the variability in perceptual evidence – although sometimes to a greater or lesser degree than predicted by an ideal observer model (Zylberberg et al., 2014, Spence et al., 2016, Boldt et al., 2017). Other work has revealed how people adjust their confidence criteria in the face of changing stimulus uncertainty (Aitchison et al., 2015, Adler & Ma, 2018, Denison et al., 2018). However, such results rely on comparing models fit across multiple trials and admit of heuristic accounts of how uncertainty affects confidence. Establishing that uncertainty on individual trials is used to inform confidence judgments has proven more difficult.

Neuroscience evidence makes a stronger case for uncertainty estimates informing confidence judgments. Kiani & Shadlen (2009) found that activity in area LIP in the monkey brain both accumulated evidence for particular choice options, and, when such activity was of intermediate strength, led to the monkeys opting out of their choice (a non-verbal marker of low certainty about either motion direction). Importantly, variability in LIP firing rates predicted the opt-out choice even when stimuli were held fixed, drawing a link between neural and behavioural markers of certainty about motion direction. Note that such activity is in a world-centred reference frame (reflecting certainty about the mapping between the stimulus and potential responses), rather than in a self-centred reference frame. However such a representation naturally supports prospective propositional confidence estimates (“How confident am I in choosing A or B, conditional on the evidence that I have gathered so far?”). The opt-out task is thus an ambiguous case – it can be solved by relying on world-centred uncertainty estimates, or self-centred (metacognitive) confidence estimates, and it is hard to tell which is in play based on behaviour or neural data alone.

Geurts et al. (2022) asked human participants to estimate the orientation of a tilted grating and judge their confidence that their estimate was accurate. Within a Bayesian framework, it was expected that the more precise the representation of orientation in visual cortex (the smaller the posterior uncertainty), the larger the propositional confidence in the tilt estimate should be. This was the case in subjects’ behaviour. The authors then used a machine learning approach to decode trial-by-trial uncertainty in the representation of particular orientations from fMRI voxel patterns within visual cortex. They found that reported confidence was negatively correlated with such a readout of uncertainty, even when the stimulus was held fixed.

There is thus good evidence that a) the brain represents uncertainty about a wide range of quantities b) that such uncertainty is used to inform metacognitive judgments. It remains unclear how and whether a similar scheme is maintained beyond sensory representations – for instance, when judging confidence in being able to remember something. Recent fMRI evidence suggests similar population-level representations of uncertainty in visual working memory (Li et al., 2021), and single unit activity in the human hippocampus predicts retrieval confidence levels (Rutishauser et al., 2015). Sampling schemes offer another potential solution, allowing probability distributions over internal states to be formed by drawing samples from internal models (Fiser et al., 2010).

4.2 Propositional confidence

Representing certainty or uncertainty in a self-centred frame of reference – what I refer to as propositional confidence – is the foundation of metacognitive judgments. Computationally, this can be achieved by transforming an internal (sensory or mnemonic) representation z into an estimate of confidence in taking an action based on z. For instance, if z indicates a probability distribution (posterior) over possible orientations (**Figure 2**) and the observer’s task is to say whether the orientation is clockwise or counterclockwise (a binary variable, d), a confidence judgment can be derived from computing $p(d = a|z, a)$ – the probability that action a picked out the correct world state d, given z. In a situation where one’s action is based solely on z, then propositional confidence is a nonlinear transformation of z. However, if there are additional sources of decisional or metacognitive noise, or if additional information arrives after committing to a decision, then propositional confidence should also be affected by these factors (Fleming & Daw, 2017). In all these cases, propositional confidence should be closely informed by estimates of uncertainty reviewed in the previous section. The upshot is a confidence estimate in the “frame of reference” of the accuracy of one’s own judgments – a self-related frame of reference.

It is natural to think of such a change in reference frame as being retrospective – I process some information, make a decision, and then reflect on whether my decision was correct. Indeed, as we will see, post-decisional processing is an

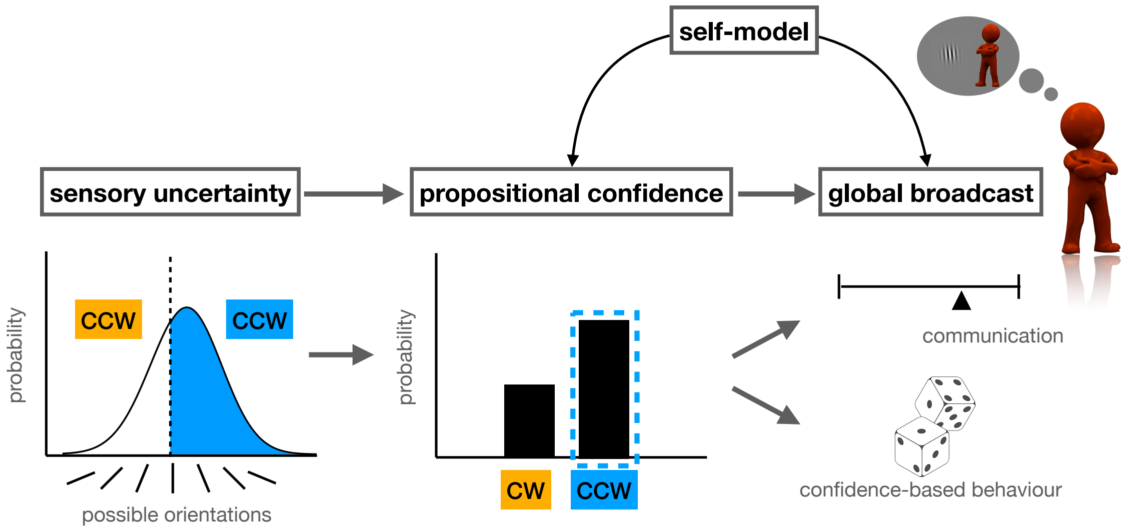

Figure 2: Graphical illustration of components of a perceptual metacognitive judgment. A generative model defines how an observer forms a belief about the state of the world – here, the orientation of the stimulus – from a noisy sensory measurement. This belief over possible orientations is associated with sensory uncertainty, and is converted into propositional confidence conditional on a categorical decision – here, whether the stimulus is tilted clockwise (CW) or counterclockwise (CCW). A propositional confidence estimate is globally broadcast for communication or usage in confidence-based behaviours, for instance guiding risk-sensitive decision-making. Background beliefs about a range of factors influencing self-performance are furnished by a self-model and influence the formation of metacognitive judgments.

important empirical signature of this change in reference frame. But propositional confidence can also be prospective. I might estimate, based on some uncertain information, the likelihood that a hypothetical decision based on that information would be correct. Such prospective judgments can apply to propositions rather than individual actions – for instance, the proposition that "I will remember this particular word" or "I will score a goal" (**Figure 1**). These prospective confidence estimates may therefore underpin classical judgments of learning, or aspects of self-confidence about ability.

More recently, defining decision confidence as a Bayesian probability of being correct has been challenged on both empirical and theoretical grounds. Empirically, confidence closely tracks the probability of making a particular choice, rather than objective notions of accuracy. For instance, if choice probability is biased by perceptual illusions, confidence often follows suit (Caziot & Mamassian, 2021, Gallagher et al., 2019). Theoretically, it is also hard to define notions of accuracy for subjective decisions, such as value-based choices or aesthetic preferences – and yet we can still evaluate confidence in such decisions (De Martino et al., 2013, Lebreton et al., 2015). Instead, a more general computational definition posits that propositional confidence reflects the probability of making a self-consistent choice across multiple presentations of the same decision problem (Caziot & Mamassian, 2021, Koriat, 2012, Boundy-Singer et al., 2023).

Computing propositional confidence

Consider a visual perceptual task in which the decision-maker should classify the orientation of a stimulus s as clockwise (CW) or counterclockwise (CCW) relative to some arbitrary boundary m (Bang & Fleming, 2018). On a single trial, the observer makes a sensory measurement X_i. The posterior over possible orientations s is then:

$$p(s|X_i) \propto p(X_i|s)p(s)$$

Because measurements are affected by noise, for a single stimulus, the measurement X_i is a bit more or a bit less than the true s. This can often be controlled by the experimenter, for instance by adjusting the contrast of a grating or the coherence of a patch of randomly moving dots. Under greater noise, the likelihood of s given X_i becomes wider (the first term on the righthand side), because the measurement is potentially consistent with a wider range of true orientations. Assuming the prior stays constant, this also leads to a more uncertain posterior over s (the lefthand side).

The observer now has an internal belief with some sensory (or mnemonic) uncertainty attached to it. But she still needs to act on this information – in this example, saying whether the orientation is CW or CCW to the boundary. Doing so requires specifying which actions are possible (mappings from s to a) and the cost or reward associated with each. Here it is useful to specify an intermediate variable that captures relevant parts of the stimulus space. d is CCW when $s < m$ and CW when $s > m$. In the case of a simple perceptual decision-making task that rewards correct decisions, the cost function $C(d, a)$ is 1 when $a = d$, and 0 otherwise. We can now define a new form of certainty about possible actions (**Figure 2**):

$$p(d_{CCW}|\hat{s}) = \int_{-\infty}^{m} p(s|X_i)$$

$$p(d_{CW}|\hat{s}) = \int_{m}^{\infty} p(s|X_i)$$

where $\hat{s}$ indicates the observer's estimate of s.

Once we have committed to a potential action (an action that will occur, or has occurred), we can use the above probabilities to compute the probability that the action was correct (that $C(d, a) = 1$):

$$p(a = d|\hat{s})$$

This quantity is what I refer to as *propositional confidence* (Pouget et al., 2016).

There have been two broad approaches to studying the behavioural and neural basis of propositional confidence. One is to simply ask for subjective reports of confidence about a future or past decision. These confidence judgments are higher for objectively correct than incorrect decisions, showing sensitivity to performance, albeit often corrupted by additional metacognitive noise (Shekhar & Rahnev, 2021). Convergent findings have emphasized the importance of the human prefrontal cortex for the fidelity of propositional confidence estimates, with a meta-analysis revealing that activity in medial and lateral prefrontal cortex, precuneus and ventral striatum covaries with judgments of confidence in memory and perceptual tasks (Vaccaro & Fleming, 2018).

A second approach harnesses statistical signatures of confidence in a self-centred (decisional) frame of reference. A prominent signature here is the "folded X" pattern: when confidence is plotted against objective measures of signal strength (the inverse of decision difficulty), propositional confidence should increase with signal strength for correct trials, and decrease with signal strength for error trials. The intuition here is that, while errors on easier trials will be less frequent, those that do occur will be accompanied by significant evidence against the chosen option, leading to lower confidence. This pattern is seen in both human and animal confidence data (Sanders et al., 2016), and has been used as a marker of confidence-related physiological and neural signals (Urai et al., 2017). In a seminal study, Kepecs and colleagues found that neurons in rodent OFC showed statistical signatures of confidence in a odour discrimination task (Kepecs et al., 2008). Confidence signatures in OFC predict confidence-related behaviour (waiting for a reward, conditional on performance) and generalise across both auditory and olfactory decisions (Masset et al., 2020), with inactivation of this brain area impairing metacognition but not performance (Lak et al., 2014). The rodent OFC is therefore a candidate neural substrate for propositional confidence.

A similar approach was adopted by Bang & Fleming (2018) in humans, in a fMRI study which manipulated both a proxy for sensory uncertainty (motion coherence) and the difficulty of the choice. Human participants viewed a random dot motion stimulus which indicated a particular direction around the circle with a given uncertainty, controlled by coherence. They then saw a decision boundary appear, before being asked to decide whether the motion direction was clockwise or counterclockwise of the boundary. This design dissociates propositional confidence in a choice (which is affected by both sensory uncertainty and decision difficulty) from sensory uncertainty (though here uncertainty was not directly assayed from neural representations, and was confounded with stimulus properties). Whereas sensory uncertainty (motion coherence) was related to activity in extrastriate visual and parietal cortex (notably, areas MT+ and bilateral intraparietal sulcus, a human homologue of LIP), signatures of propositional confidence were instead observed in ventromedial prefrontal cortex (perigenual anterior cingulate cortex, pgACC).

A complementary perspective on the neural basis of propositional confidence is provided by the literature on error monitoring, which has typically used speeded response-conflict tasks to induce response errors under time pressure. A canonical finding is that posterior medial frontal cortex neurons covary with error commission in the absence of feedback, generating an error-related negativity at the scalp surface (Desender et al., 2021). The ERN peaks approximately 100ms after the erroneous action, and arises before any feedback is given about the accuracy of the response. In animal models, post-decisional firing rates of neurons in the prefrontal frontal cortex and dopaminergic midbrain have also been shown to covary with choice correctness before explicit feedback is given (Tsujimoto et al., 2010, Middlebrooks & Sommer, 2012, Kepecs et al., 2008). Within a reinforcement learning framework, one perspective on such signals is that they reflect proxies for reward prediction errors driven not by external feedback, but by internal levels of choice confidence (Guggenmos et al., 2016, Lak et al., 2017).

More recently, it has been argued that post-decisional accumulation of evidence facilitates the formation of propositional confidence (Desender et al., 2021). The idea here builds on classical evidence accumulation frameworks that posit samples of sensory information are accumulated over a few hundred milliseconds before hitting the bound for one or other choice option. Such models have been highly successful in accounting for choice and response time behaviour in a variety of decision scenarios, and neural correlates of evidence accumulation signals have been identified in humans and animals. Moreover, as we saw above, the dynamics of evidence accumulation within the choice period provide a neural representation of uncertainty that can be used to inform confidence (Kiani & Shadlen, 2009). Pleskac & Busemeyer (2010) additionally proposed that this evidence accumulation process may continue after a decision has been made, to inform estimates of decision confidence and potentially leading to changes of mind (van den Berg et al., 2016a, Resulaj et al., 2009).

Post-decisional processes may either continue to accumulate sensory evidence for and against available choice alternatives (world-centred reference frame), or accumulate evidence about the accuracy of the preceding choice (self-centred reference frame). Murphy et al. (2015) and colleagues found that the ramping-like characteristics of a centroparietal EEG signal, the Pe, was consistent with post-decisional evidence accumulation in a self-related reference frame. The post-decisional build-up rate of this signal was proportional to the speed of subjective error detection, and reached a constant amplitude at the point of detection that was independent of error-detection RT. Interestingly, the Pe signature is similar to the centroparietal positivity (CPP) that has been linked to pre-decisional evidence accumulation in a world-centred reference frame. This suggests that the CPP and the Pe may reflect a general evidence accumulation circuit that can flexibly adapt reference frames in the service of both first-order performance and metacognition. Boldt & Yeung (2015) found that the Pe amplitude also predicts graded ratings of confidence in choice, highlighting how this accumulation signal goes beyond all-or-nothing error detection.

These studies investigated endogenous post-decisional accumulation of evidence. It is also possible to experimentally manipulate the availability of post-decisional information. Computationally, injecting additional post-decision evidence should promote the folded X-pattern in confidence ratings, due to a greater opportunity for gaining evidence against an incorrect decision. In a study of random dot motion discrimination, providing stronger post-decision evidence indeed led to a stronger folded-X pattern in confidence ratings (Fleming et al., 2018). This folded-X signature was observed in the fMRI activity of the posterior medial frontal cortex, consistent with this region (negatively) accumulating evidence in a frame of reference of choice accuracy, and providing a computational bridge between studies on confidence and error monitoring.

4.3 Global broadcast and communication

For propositional confidence to be useful to guide flexible behaviour, it should be broadcast to a number of different consumer systems (Baars, 1993). This would allow different propositional confidences to be compared in a common frame of reference – allowing the agent to decide, for instance, that they are more likely to be successful in judgments of one or other task or sensory modality (Aguilar-Lleyda & de Gardelle, 2021). The global broadcast of confidence can also be used as a learning signal in lieu of external feedback – allowing the online detection of errors, and consequent adjustments to behaviour (Guggenmos et al., 2016). Interestingly, propositional confidence may emerge in parallel to the decision (or proposition) itself, and be used to shape the ongoing decision process, for instance, controlling the termination of evidence accumulation (Balsdon et al., 2020) or guiding the next step in a sequential decision (van den Berg et al., 2016b). Finally, global broadcast of propositional confidence is important for the public sharing of metacognitive representations in group settings: we might say to a colleague, “I believe this is the right thing to do”, thereby influencing the course of the group’s decision (Bahrami et al., 2010, Shea et al., 2014). Mappings between “private” feelings of confidence and public utterances lead to additional computational considerations. In a collaborative context, it is important to align the distribution of our confidence statements with those of others, to avoid dominating a group interaction (or being dominated ourselves; Bang et al., 2017). However, if we wish to strategically influence the group, it might be advantageous to over (or under-)state our public confidence (Hertz et al., 2017).

Global broadcast is proposed to covary with conscious awareness of a range of mental content, including metacognitive representations (Dehaene et al., 2017). This implies that forms of propositional confidence that remain restricted to a particular sensorimotor pathway, and not globally shared, may underpin non-conscious forms of metacognition (Charles et al., 2013, Logan & Crump, 2010). We may also consciously experience other forms of perceptual uncertainty beyond propositional confidence (Morrison, 2016), and such uncertainty estimates may themselves affect what content is globally broadcast (Shea & Frith, 2019).

Behaviorally, elegant work has shown that people are able to estimate and compare propositional confidence about decisions made in two different sensory modalities, indicating that domain-specific confidence estimates can be broadcast and shared (de Gardelle et al., 2016). There is also emerging evidence that metacognitive capacity (measured as the noise in metacognitive judgments, relative to performance) covaries across perceptual and cognitive tasks, suggesting a global resource that is leveraged to monitor self-performance (Rouault et al., 2018b, Mazancieux et al., 2020, Boundy-Singer et al., 2023).

A common currency for confidence may be supported by modality-independent confidence signals in the rodent (Masset et al., 2020) and human (Morales et al., 2018) prefrontal cortex. Recently, an impressive study conducted single-unit recordings in human neurosurgical patients performing two distinct tasks in which errors were relatively common (Fu et al., 2022). At the population level, posterior medial frontal cortex cells formed a high-dimensional representation that allowed simple linear decoders to read out both domain-general error signals, and simultaneously to differentiate domain-specific aspects of performance monitoring, such as the task and type of response conflict that gave rise to the error.

Performance monitoring signals are sensitive not only to the objective act of making an error, but also to subjective error awareness (Nieuwenhuis et al., 2001) and decision confidence (Boldt & Yeung, 2015), albeit with some intriguing dissociations that may indicate specific roles in global broadcast. The Pe (described in the last section as being a candidate for post-decisional evidence accumulation) has been linked to error awareness and shown to covary with subjective confidence, whereas the ERN and its pMFC source are thought to also operate unconsciously (Charles et al., 2013). Consistent with this perspective, fMRI neural correlates of evidence against a choice were tracked in pMFC (the neural generator of the ERN), whereas more anterior prefrontal regions covaried with subjective confidence (Fleming et al., 2018).

An alternative perspective on the neural basis for broadcast and communication is provided by studies that have explicitly manipulated the requirement for a metacognitive judgment. For instance, one might compare trials on which a decision is made together with a metacognitive judgment of confidence, against a control condition where the same kind of decision is made, but now the rating is about another property of the stimulus (eg its brightness or size). Such comparisons have highlighted a network of prefrontal regions, notably the dorsal anterior cingulate cortex and lateral frontopolar cortex, in which activity is heightened when metacognitive judgments are required (Fleming et al., 2012, Qiu et al., 2018, Yeon et al., 2020).

A particularly detailed perspective on metacognitive judgment-related neural activation was provided by Gherman & Philiastides (2018). Using EEG-informed fMRI, they could separate early neural activations correlating with confidence, from later activations linked to the requirement for an explicit metacognitive judgment. Early confidence related signals were seen in the ventromedial prefrontal cortex (in a similar pgACC region identified by Bang & Fleming), whereas later judgment-related activation was seen in the lateral frontopolar cortex. Finally, in the study by Geurts et al. (2022) described above, the decoder’s readout of sensory uncertainty in early visual areas was correlated with univariate fMRI signals in the prefrontal cortex, consistent with domain-specific uncertainty estimates informing globally available estimates of propositional confidence.

An alternative approach to assaying the behavioural and neural signatures of broadcast and communication experimentally dissociates “private” estimates of propositional confidence from the “public” estimates that are communicated to others. One natural way of achieving this is in a group context where individuals have to pool their confidence estimates to drive a group decision. Previous work has shown that when individuals are collaborating in this way, the two partners rapidly and naturally adapt their confidence levels to converge on a common scale, so that one does not dominate the other (Bang et al., 2017). In an fMRI study of such social coordination about random dot motion judgments, it was found that whereas ventromedial prefrontal cortex (pgACC) covaried with private estimates of propositional confidence, as in previous work, lateral frontopolar cortex additionally carried information about the extent to which a private-public mapping should be adjusted when communicating a public judgment (Bang et al., 2020). These findings are intriguing in light of other work emphasizing the role of frontopolar cortex in metacognitive efficiency (Fleming et al., 2010, Miyamoto et al., 2018, Baird et al., 2013, McCurdy et al., 2013, Allen et al., 2017). Such findings have often been interpreted as indicating a role for the frontopolar cortex in supporting metarepresentations, with the impairment of the functions of this region leading to greater metacognitive noise. An alternative hypothesis is that the frontopolar cortex constrains metacognitive efficiency by maintaining a stable private-public mapping, with instability in this mapping manifesting as a weaker coupling between metacognition and performance.

4.4 The role of self-models

Up until now we have considered a relatively lean, minimal notion of propositional confidence, one that is directly informed by the internal states driving behaviour (sometimes known as a “first-order” model of confidence formation). However there is also a range of findings on human metacognition that suggest propositional confidence makes use of a richer (implicit) model of the factors affecting performance. The idea here is that, just as we build up a theory of how other minds work, we also build up a model of the factors affecting our own mental operations, and bring that model to bear when making metacognitive judgments (Nelson & Narens, 1990). Some of these background beliefs about how our minds work may be acquired via learning, or culturally inherited – as when children are instructed that feelings of fluency might produce misleading boosts in confidence, and they would be wise to slow down and reconsider their answer (Heyes et al., 2020). Differences between cultures in how these beliefs are acquired may account for findings of cultural differences in confidence and metacognition (Yates et al., 1998, van der Plas et al., 2022), and how people process self-related feedback (Kitayama et al., 1997). Other background beliefs may be more innate and furnished by evolution, such as associations between interoceptive states and confidence (Allen et al., 2016, Fiacconi et al., 2016).

A long-standing proposal is that model-based contributions to metacognition rely on extensions of the model(s) that guide our predictions of the mental states and behaviours of other people – a capacity known as theory of mind or mentalizing (Carruthers, 2009). More generally, the implication is that we do not have direct access to first-order cognitive processes, and instead have to infer their status from a variety of cues, just as we infer what others think or feel from observing their behaviour. This view casts (model-based) metacognition as operating on similar principles to perception, in that both rely on the principles of (unconscious) inference.

A prominent theory in the metamemory literature proposes that a variety of cues affect metacognitive judgments via an inferential process. This renders metacognition susceptible to illusions and distortions – metacognitive analogues of perceptual illusions (Alter & Oppenheimer, 2009). For instance, we may hold a belief that faster decisions are more likely to be accurate, and use these feelings of fluency to inform our confidence estimates (Kiani et al., 2014). Similar boosts in fluency can be achieved by increasing the brightness of a face stimulus (Busey et al., 2000), or the font size of a word stimulus (Hu et al., 2015) – leading to greater confidence in recall without any change in performance. Other work indicates that interoceptive factors influence confidence judgments even if they are irrelevant to the decision at hand (Fiacconi et al., 2016). For instance, Allen et al. (2016) found that subliminally presented disgusted faces led not only to changes in pupil dilation and heart rate, but also modulated confidence in a perceptual (random dot motion) decision. The existence of these effects indicate the influence of an (implicit) self-model at work in the construction of explicit confidence judgments in a range of domains.

There has been relatively little work assaying the computational basis of model-based metacognitive inference, or how such models are instantiated in the brain. One possibility is that self-models furnish beliefs about the parameters of the confidence formation process (which may not always match the actual parameters of such a process; Fleming & Daw, 2017, Khalvati et al., 2021, Marcke et al., 2022). For instance, Hu et al. (2021) suggested that people’s judgments of learning are constructed by integrating their processing experience on single trials with prior beliefs about how different cues affect memory performance – even if such cues do not promote objective success. Similarly, in the perceptual domain, Winter & Peters (2022) found that people misperceive sensory noise in the periphery of the visual field, leading to an inflation of perceptual confidence relative to perceptual acuity. This work implies a close connection between model-based influences on metacognition and the role of priors in propositional confidence formation. In an elegant experiment, Marcke et al. (2022) modulated people’s priors on perceptual confidence through the use of false feedback on their relative scores compared to other participants. This influence was best-captured by a model in which the parameters relating evidence accumulation to confidence were modified by a prior belief, without affecting objective accuracy or response times.

Effects of self-action on metacognitive judgments are another potential manifestation of model-based influences on confidence formation. Fleming & Daw (2017) proposed that a confidence computation may leverage information provided by one’s own actions when inferring whether a decision is likely to be correct. Tell-tale signs of this effect have been confirmed empirically: metacognitive sensitivity is often better when confidence judgments are provided after, compared to before, an explicit decision has been made (Siedlecka et al., 2016, Pereira et al., 2020, Wokke et al., 2020), with activity in frontopolar and insula cortex, and beta-band synchrony between motor and frontal cortex, hypothesized to mediate the impact of self-actions on metacognitive estimates (Wokke et al., 2020, Pereira et al., 2020). Conversely, metacognitive sensitivity is reduced when a task-relevant motor action is disrupted by applying transcranial magnetic stimulation over premotor cortex (Fleming et al., 2015). While these findings remain to be fully assimilated into computational models of confidence, they indicate that metacognitive judgments are sensitive to a range of internal cues that go beyond first-order performance.

More broadly, as noted above, one influential view is that model-based influences on metacognition may draw on similar resources to those supporting mentalizing about others. There is circumstantial evidence for this link, with similar developmental trajectories (both metacognition and mentalizing emerge around the age of 3-4) and overlap in neural correlates (particularly in the medial prefrontal cortex; Vaccaro & Fleming, 2018). Recently, in an elegant series of studies, Nicholson and colleagues found that perceptual metacognitive sensitivity on a task requiring explicit confidence judgments (but not, intriguingly, one requiring an implicit gamble of the kind often used in animal metacognition experiments) correlated with mentalizing abilities, and was impaired in subjects with autism spectrum disorder. In addition, a secondary mentalizing task (but not another, equivalently demanding task) interfered with explicit metacognitive judgments (Nicholson et al., 2021). Together this work suggests that the model-based component of human metacognition may co-opt social cognitive resources – although how such resources interface with the bottom-up aspects of propositional confidence formation reviewed above remains to be determined.

5 CONFIDENCE FORMATION AND THE PSYCHOLOGY OF METACOGNITION

We can now take stock and consider how these different computational stages interact, and map onto the psychology of metacognition. First, myriad uncertainties exist at all stages of perception and cognition. Such uncertainties encompass not only well-studied perceptual systems, but also internal uncertainties arising from memory, or uncertainties in interoception. Sensitivity to uncertainty is a central aspect of (first-order) Bayesian computation, but alone is not evidence for metacognition. A further stage encodes confidence relative to a proposition – a (hypothetical) statement or decision – in a self-centred reference frame. This stage qualifies as (model-free) metacognition in that it has a mental state of the self (the proposition) among its correctness conditions (Carruthers & Williams, 2022). A sensible agent will make use of domain-specific uncertainty when forming propositional confidence, and in some constrained scenarios the latter will be a minor transformation of the former (consider, for instance, a posterior belief over potential motion directions that is transformed into propositional confidence in a specific choice option). It is therefore important to be aware that some tasks held to measure metacognition such as the opt-out task are often ambiguous with respect to whether they are tracking metacognitive (propositional) confidence, or world-centred uncertainty. Propositional confidence can be globally broadcast and used in a range of metacognitive control functions – including strategically adjusting how confidence estimates are communicated to others. Finally, the formation of propositional confidence may itself be influenced by an implicit model of how first-order cognitive systems operate.

These different stages can tentatively be mapped to systems-level interactions between brain areas. As noted above, early sensory areas may represent uncertainty over sensory variables such as motion direction, whereas prefrontal regions such as pMFC and vmPFC track propositional confidence. Lateral frontopolar cortex is recruited to allow global broadcast and strategic communication of propositional confidence estimates. We can also make a tentative proposal that regions involved in theory of mind – including ventromedial and dorsomedial PFC, and temporoparietal junction – may support self-models that contribute to model-based metacognition (Wittmann et al., 2016, Vaccaro & Fleming, 2018). However, it is important here to distinguish between propositional confidence in self-actions, propositional confidence in the actions of others, and the roles that models of self and other play in the formation of metacognitive judgments. Recent brain imaging studies suggest propositional confidence formation in self and other draws on distinct brain networks (Bang et al., 2022, Jiang et al., 2022), but that dorsomedial PFC may act as a common node for furnishing model-based information for both metacognition and mentalizing (Jiang et al., 2022, Wittmann et al., 2016).

Although these components have up until now been presented as distinct, this is for didactic convenience, and we should expect mutual interactions between each to be the norm. Indeed, understanding the interrelationships between different stages of metacognitive computation is only just beginning to be investigated (Geurts et al., 2022, Shekhar & Rahnev, 2018, Bang & Fleming, 2018), but represents a major goal for the field (Rahnev et al., 2022). One possibility is that neural codes within different frames of reference emerge and are maintained in parallel, serving different computational goals. For instance, evidence may be accumulated about particular sensory features (world-centred frame of reference), and simultaneously about (future or past) choice correctness (self-centred frame of reference) – with the latter feeding back to set the bound on current or future sensory evidence accumulation (Balsdon et al., 2020). As such, it is likely to be more fruitful to view metacognition as emerging from a set of dynamically interacting internal states, some of which are world-centred, and others of which encode beliefs about one's propositions or decisions (Yeung & Summerfield, 2012).

6 REVISITING CURRENT CONTROVERSIES

6.1 Biases and suboptimalities in confidence

A long-running debate in the field is between those who consider confidence (and by implication, metacognition) to be inherent to the decision process, and those who consider it to depend on additional machinery or computation. In emphasizing multiple computational components, the current framework provides a resolution of this tension. In certain scenarios, such as the decision to opt out of a well-constrained decision problem, propositional confidence can be derived from a direct transformation of the accumulated evidence for one or other choice (Kiani & Shadlen, 2009, van den Berg et al., 2016a). But in other scenarios, particularly when post-decisional accumulation of evidence is in play, or when there are multiple model-based cues to confidence, a second stage of propositional confidence formation may be involved – particularly when it is functionally advantageous to broadcast such confidence to multiple distinct consumer systems. In that situation, dedicated machinery for the readout and usage of propositional confidence (for instance in PFC) may be the norm rather than the exception, with lesions or damage to these downstream areas manifesting as selective metacognitive deficits, and presenting more opportunity for deviations from ideal observer models to occur.

A fruitful approach to pursuing the computational basis of metacognition, then, is to explicitly model these different stages and ask how noise or suboptimalities within each component may contribute to metacognitive inefficiencies (Shekhar & Rahnev, 2021, Guggenmos, 2022, Mamassian & de Gardelle, 2022). For instance, Boundy-Singer et al. (2023) identify meta-uncertainty about sensory uncertainty as a key domain-general constraint on the fidelity of propositional confidence estimates in both perceptual and cognitive decision tasks. In turn, constraints on post-decisional evidence accumulation may affect the extent to which confidence estimates faithfully track performance (Pleskac & Busemeyer, 2010, Desender et al., 2022). When interacting with others, there is a requirement to maintain distinct models for ourselves and others, and selecting the correct model may be computationally demanding: Wittmann et al. (2016) found that when tracking the performance of oneself and others, people sometimes "merged" their feedback with those of others. This intertwining of models of self- and other-performance was associated with differences in activity in dorsomedial prefrontal cortex, and disrupting this area using TMS led to greater self-other mergence (Wittmann et al., 2021) – suggesting that one function of this brain region is not only to support models of ourselves and others, but also to keep these models apart. More generally, different suboptimalities may co-exist and the same kind of computational constraints that affect first-order cognition are likely to affect the suboptimality of metacognition (Rahnev & Denison, 2018).

One metacognitive bias that has received particularly detailed theoretical and empirical scrutiny is the "positive evidence" bias (PEB). The PEB manifests as confidence being more affected by evidence in favour of a choice than evidence against it (Zylberberg et al., 2012), such that an increase in overall evidence results in boosts in confidence despite performance remaining unaffected. Initial theoretical explanations proposed that the PEB may result from a bias in the broadcast or readout of propositional confidence estimates, or a heuristic applied to evidence spaces that are often detection- rather than discrimination-like (Miyoshi & Lau, 2020, Maniscalco et al., 2021). More recently, though, empirical and modeling studies have led to surprising conclusions that constrain the origins of the PEB. First, a PEB emerges within a convolutional neural network that is trained to both discriminate digits and estimate the probability correct of these classifications – indicating that a PEB may not be a foible of human metacognition, but a core feature of how high-dimensional evidence spaces are mapped to propositional confidence (Webb et al., 2022). Second, the PEB can be "flipped" – creating a negative evidence bias – if the decision is reframed as a search for the weaker response option (e.g. fewer dots, or a disliked item; Sepulveda et al., 2020). Together these findings point towards a model in which the PEB may be a feature of how propositional confidence is formed, rather than a bias in the tracking of domain-specific uncertainties.

6.2 Sources of domain-generality

Another contested issue is the extent to which metacognitive capacities should be considered domain-general, or domain-specific. Behaviourally, individual differences in metacognitive efficiency have been shown to be correlated across distinct task domains, after controlling for correlations in performance (Mazancieux et al., 2020, Rouault et al., 2018b, Ais et al., 2016, Faivre et al., 2018). However, the strength of these correlations is often weak and variable, especially in smaller samples used in neuroimaging research (see Rouault et al., 2018b, for a meta-analysis). In addition, there are concerns that factors only indirectly related to metacognitive capacity may contribute to findings of domain-generality – such as how confidence scales are used, or the adoption of a particular thresholds for post-decisional evidence accumulation (Xue et al., 2021, Desender et al., 2022). Findings of domain-generality in metacognitive bias (average confidence level) are more robust, and have been related to features of both personality and mental health (see Sidebar).

Set against findings of correlated individual differences are findings of both domain-specificity in the neural basis of metacognition, and domain-specific impairments in metacognitive efficiency following lesions or experimental intervention. One particularly consistent set of findings points to a selective role for medial parietal cortex (precuneus) in metamemory (McCurdy et al., 2013, Baird et al., 2013). Accordingly, lesions to the frontopolar cortex (but not precuneus) lead to impaired metaperceptual efficiency, but leave metamnemonic efficiency (as assayed by recognition memory confidence) intact (Fleming et al., 2014). The reverse dissociation is seen with theta burst TMS to the precuneus, which impairs metamnemonic but not metaperceptual efficiency (Ye et al., 2018).

The current framework provides an opportunity to integrate these findings. A positive manifold in individual differences in healthy metacognition may be mediated by common downstream processes involved in the formation of propositional confidence and/or its global broadcast. Conversely, domain-specific limitations may be imposed by how domain-specific uncertainty is propagated into a propositional confidence computation, and/or the fidelity of model-based estimates of uncertainty parameters (uncertainty about uncertainty; Boundy-Singer et al., 2023, Khalvati et al., 2021). For instance, one plausible although speculative role for the precuneus in metamemory is that it is involved in translating uncertainty-related information carried by hippocampal neurons (Rutishauser et al., 2015) into a (prospective or retrospective) propositional confidence judgment. Understanding these interactions will be aided by new data analysis approaches that seek to understand which variables can be easily read out from mixed selectivity neural populations – with the possibility that both domain-general and domain-specific components of confidence formation co-exist within the same brain area (Fu et al., 2022, Morales et al., 2018). At a behavioural level, future work should seek to move beyond examining correlations in descriptive statistics such as meta-*d'*, and instead seek to characterize the computational stages at which domain-generality in metacognition emerges (Boundy-Singer et al., 2023, West et al., 2022).

Individual and group differences

Metacognitive efficiency shows moderate test-retest reliability, both across different sessions of the same experiment (Fleming et al., 2010, Ais et al., 2016) and across different days (Wright et al., 2012). Metacognitive bias (calibration) shows stronger test-retest reliability, with stable confidence "fingerprints" seen across different tasks and testing sessions (Ais et al., 2016). A number of studies have linked local and global metacognitive biases to individual differences in transdiagnostic mental health symptoms, including anxiety, depression, self-esteem and compulsivity (Hoven et al., 2019, Seow et al., 2021). Conversely, metacognitive efficiency is predictive of individual differences in dogmatism about real-world issues such as politics and climate change (Rollwage et al., 2018, Fischer et al., 2019), with parameters governing confidence formation correlating with people's openness to new information (Schulz et al., 2020).

7 WHERE NEXT?

7.1 Searching for common computational principles

As indicated in the preceding section, a key next step is to move beyond the useful but artificial division of metacognition research into distinct domains, such as memory or perception, to characterize common computational principles that constrain metacognitive capacity. In this respect, the often segregated fields of perceptual and memory metacognition research can learn a lot from each other, and further cross-fertilisation will no doubt reap benefits. For instance, the metaperception field has developed psychophysical paradigms that allow the detailed computational modeling of pre- and post-decisional processes, in which hundreds of trials per participant are often required to fully characterize the joint distribution of accuracy, response time and confidence. These endeavours have been accelerated by the development of the Confidence Database and the adoption of consensus goals in the field (Rahnev et al., 2020, 2022). Conversely, the metamemory field has tended to leverage more naturalistic stimuli (memory for faces, for instance) and developed clever experimental designs to carefully unpack the contribution that a range of cues make to metacognitive judgments – for instance, the illusory boosts in confidence that ensue from manipulations of fluency.

7.2 From local to global metacognition

Most research on propositional confidence has focused on "local" judgments of performance on individual trials or task episodes. In contrast, a distinct literature in social and clinical psychology has focused on how people evaluate themselves at a global level – for instance, their self-efficacy, or estimates of their abilities relative to others. These global self-evaluations are related to future attainment (via an impact on motivation and task engagement) and may govern adaptive behaviour such knowing when to seek help or offload to the environment. However, little is known about how local metacognitive computations influence and shape self-evaluations over this longer timescale.

One fruitful approach considers global confidence as a higher-order prior on estimates of (local) propositional confidence (Marcke et al., 2022, Boldt et al., 2019), which can be naturally modeled as a probability distribution over expected success (Rouault et al., 2019). In the absence of any local task experience, people access this prior when making confidence judgments – for instance, estimating the chances they will score from a free kick (**Figure 1**). In turn, this prior can be updated in light of local (retrospective) confidence in individual actions or decisions. Tentative evidence for this view comes from experiments in which subjects provided intermittent global confidence estimates on a perceptual task (Rouault et al., 2019, Lee et al., 2021). Global confidence was informed by local confidence fluctuations during the previous block, and using fMRI, it was found that the vmPFC and precuneus may integrate local confidence over longer timescales to track aggregate self-performance (Rouault & Fleming, 2020, Wittmann et al., 2016).

Another perspective on how propositional confidence unfolds over longer timescales is provided by studies that have examined how subjects estimate the probability of making task errors based on recent experience. In an elegant paradigm, Purcell & Kiani (2016) found that subjects track a prior on propositional confidence by integrating evidence over multiple trials, and leverage this prior on expected task accuracy to decide whether to switch strategy (in effect, reaching a threshold at which they decide to blame the error on the task, rather than themselves). Neurons in monkey pMFC (anterior cingulate cortex) were found to integrate information about previous trials and drive decisions about whether to switch strategy (Sarafyazd & Jazayeri, 2019). Similarly, in human pMFC, neuronal populations signal expected conflict probability (a proxy for propositional confidence) across trials as a state variable that is orthogonal to within-trial dynamics, just as might be expected for neural activity encoding a prior on confidence level (Fu et al., 2022).

7.3 Opportunities for metacognitive interventions

Interventions to modify metacognition are in their infancy, but here too progress could benefit from understanding which computational stage(s) are being affected. Previous work has suggested that metacognitive efficiency may be modulated in response to meditation (Baird et al., 2014), drugs (Hauser et al., 2017), neurofeedback (Cortese et al., 2016), brain stimulation (Shekhar & Rahnev, 2018) and training (Carpenter et al., 2019). However, with some notable exceptions (Shekhar & Rahnev, 2018), the locus of action of these effects remains poorly understood. Knowing which steps in a computational chain are targeted by an intervention helps both to identify how and whether metacognitive boosts are likely to generalise beyond the lab, and what functional benefits they might provide. For instance, a beta blocker may inhibit the contribution of model-based interoceptive cues to confidence estimates, therefore improving metacognitive efficiency on a constrained laboratory task, but impairing it in situations where those cues are more valid. As another example, delivering feedback to improve confidence calibration over a period of two weeks shows promise in elevating metacognitive efficiency, not only on the trained task, but also more broadly (Carpenter et al., 2019). However, recent work suggests that the incentives underpinning this intervention primarily acted upon the way that confidence was communicated via a confidence scale (Rouy et al., 2022) – perhaps at the level of a private-public mapping, rather than at the level of propositional confidence formation. Such an intervention may still be useful in social situations where public confidence estimates are being pooled across observers, but less useful in cases in which propositional confidence is being used for intrapersonal control.

8 CONCLUSIONS

The fields of metacognition and confidence research are natural allies, but have often been uneasy bedfellows. Here I argue that metacognition research *is* the study of propositional confidence in all its forms. Once this is recognised, it opens up the problem of how different computational components of confidence formation interact – including those supporting the rich self-models that humans bring to bear when evaluating their behaviours and internal states. In this endeavour, the different subfields of metacognition research have a lot to learn from each other. There is no reason to think that representations of uncertainty are any less relevant for understanding metamemory, or that the contribution of self-models and other heuristics are no less relevant for understanding perceptual confidence. A research program that bridges this divide, and that seeks to understand the full range of computational stages underpinning human metacognition, will likely benefit from the lessons that can be gleaned from both of these literatures. In turn, disparate findings on the neural basis of uncertainty and performance monitoring can be integrated into a common framework, and a new understanding of the locus of action of metacognitive interventions achieved.

Summary points

1. Confidence research has focused on subpersonal representations of confidence and uncertainty in sensory or motor tasks.
2. Metacognition research is concerned with personal-level beliefs about performance.
3. These viewpoints can be reconciled by recognising metacognitive judgments as propositional confidence estimates about (hypothetical) decisions or actions.
4. A key step in forming propositional confidence is shifting between encoding uncertainty in world- and self-centred frames of reference.
5. Propositional confidence can be globally broadcast to support a range of metacognitive control functions, including social communication.
6. Model-based influences on confidence formation (such as beliefs and priors about performance) may share parallels with theory of mind.

Future issues

1. What are the common computational principles that constrain metacognitive capacity across domains?
2. How can computational models of confidence formation capture model-based influences on metacognition?
3. Do model-based influences on metacognition share neural and computational resources with theory of mind?
4. Can models of confidence developed in psychophysical experiments be generalised to naturalistic scenarios?
5. Can novel metacognitive interventions be developed based on a refined understanding of the computational components of confidence?

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

SMF is a CIFAR Fellow in the Brain, Mind & Consciousness Program, and funded by a Wellcome/Royal Society Sir Henry Dale Fellowship (206648/Z/17/Z) and a Philip Leverhulme Prize from the Leverhulme Trust. The Wellcome Centre for Human Neuroimaging is supported by core funding from the Wellcome Trust (203147/Z/16/Z). The Max Planck UCL Centre is a joint initiative supported by UCL and the Max Planck Society. I thank Shinobu Kitayama and John Dorsch for helpful comments on an earlier draft of this paper.

References

Ackerman R, Thompson VA. 2017. Meta-Reasoning: Monitoring and Control of Thinking and Reasoning. *Trends in Cognitive Sciences* 21(8):607–617

Adler WT, Ma WJ. 2018. Comparing Bayesian and non-Bayesian accounts of human confidence reports. *PLOS Computational Biology* 14(11):e1006572

Aguilar-Lleyda D, de Gardelle V. 2021. Confidence guides priority between forthcoming tasks. *Scientific Reports* 11(1):18320

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

Aitchison L, Bang D, Bahrami B, Latham PE. 2015. Doubly Bayesian Analysis of Confidence in Perceptual Decision-Making. *PLoS computational biology* 11(10):e1004519

Allen M, Frank D, Schwarzkopf DS, Fardo F, Winston JS, et al. 2016. Unexpected arousal modulates the influence of sensory noise on confidence. *eLife* 5:403

Allen M, Glen JC, Müllensiefen D, Schwarzkopf DS, Fardo F, et al. 2017. Metacognitive ability correlates with hippocampal and prefrontal microstructure. *NeuroImage* 149:415–423

Alter AL, Oppenheimer DM. 2009. Uniting the Tribes of Fluency to Form a Metacognitive Nation. *Personality and Social Psychology Review* 13(3):219–235

Baars BJ. 1993. A cognitive theory of consciousness. Cambridge University Press

Bahrami B, Olsen K, Latham PE, Roepstorff A, Rees G, Frith CD. 2010. Optimally interacting minds. *Science* 329(5995):1081–1085

Baird B, Cieslak M, Smallwood J, Grafton ST, Schooler JW. 2015. Regional white matter variation associated with domain-specific metacognitive accuracy. *Journal of Cognitive Neuroscience* 27(3):440–452

Baird B, Mrazek MD, Phillips DT, Schooler JW. 2014. Domain-specific enhancement of metacognitive ability following meditation training. *Journal of Experimental Psychology. General* 143(5):1972–1979

Baird B, Smallwood J, Gorgolewski KJ, Margulies DS. 2013. Medial and Lateral Networks in Anterior Prefrontal Cortex Support Metacognitive Ability for Memory and Perception. *Journal of Neuroscience* 33(42):16657–16665

Balsdon T, Wyart V, Mamassian P. 2020. Confidence controls perceptual evidence accumulation. *Nature Communications* 11(1):1753

Bang D, Aitchison L, Moran R, Herce Castanon S, Rafiee B, et al. 2017. Confidence matching in group decision-making. *Nature Human Behaviour* 1(6):1–7

Bang D, Ershadmanesh S, Nili H, Fleming SM. 2020. Private–public mappings in human prefrontal cortex. *eLife* 9:e56477

Bang D, Fleming SM. 2018. Distinct encoding of decision confidence in human medial prefrontal cortex. *Proceedings of the national academy of sciences* 115(23):6082–6087

Bang D, Moran R, Daw ND, Fleming SM. 2022. Neurocomputational mechanisms of confidence in self and others. *Nature Communications* 13(1):4238

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):1520–1531

Boldt A, Schiffer AM, 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

Boundy-Singer ZM, Ziemba CM, Goris RLT. 2023. Confidence reflects a noisy decision reliability estimate. *Nature Human Behaviour* 7(1):142–154

Busey TA, Tunnicliff J, Loftus GR, Loftus EF. 2000. Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review* 7(1):26–48

Carpenter J, Sherman MT, Kievit RA, Seth AK, Lau H, Fleming SM. 2019. Domain-general enhancements of metacognitive ability through adaptive training. *Journal of Experimental Psychology. General* 148(1):51–64

Carruthers P. 2009. How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences* 32(02):121–138

Carruthers P, Williams DM. 2022. Model-free metacognition. *Cognition* 225:105117

Caziot B, Mamassian P. 2021. Perceptual confidence judgments reflect self-consistency. *Journal of Vision* 21(12):8

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

Cortese A, Amano K, Koizumi A, Kawato M, Lau HC. 2016. Multivoxel neurofeedback selectively modulates confidence without changing perceptual performance. *Nature Communications* 7:13669

Dayan P. 2022. Metacognitive Information Theory. PsyArXiv

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* 11(1):e0147901

De Martino B, Fleming SM, Garrett N, Dolan RJ. 2013. Confidence in value-based choice. *Nature Neuroscience* 16(1):105–110

Dehaene S, Lau HC, Kouider S. 2017. What is consciousness, and could machines have it? *Science* 358(6362):486–492

Denison RN, Adler WT, Carrasco M, Ma WJ. 2018. Humans incorporate attention-dependent uncertainty into perceptual decisions and confidence. *Proceedings of the National Academy of Sciences of the United States of America* 115(43):11090–11095

Desender K, Ridderinkhof KR, Murphy PR. 2021. Understanding neural signals of post-decisional performance monitoring: An integrative review. *eLife* 10:e67556

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

Ernst MO, Banks MS. 2002. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415(6870):429–433

Faivre N, Filevich E, Solovey G, Kühn S, Blanke O. 2018. Behavioral, Modeling, and Electrophysiological Evidence for Supramodality in Human Metacognition. *The Journal of Neuroscience* 38(2):263–277

Fiacconi CM, Peter EL, Owais S, Köhler S. 2016. Knowing by heart: Visceral feedback shapes recognition memory judgments. *Journal of Experimental Psychology. General* 145(5):559–572

Fischer H, Amelung D, Said N. 2019. The accuracy of German citizens' confidence in their climate change knowledge. *Nature Climate Change* 9(10):776–780

Fiser J, Berkes P, Orbán G, Lengyel M. 2010. Statistically optimal perception and learning: From behavior to neural representations. *Trends in Cognitive Sciences* 14(3):119–130

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

Fleming SM, Daw ND. 2017. Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review* 124(1):91–114

Fleming SM, Dolan RJ. 2012. The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B-Biological Sciences* 367(1594):1338–1349

Fleming SM, Huijgen J, Dolan RJ. 2012. Prefrontal Contributions to Metacognition in Perceptual Decision Making. *Journal of Neuroscience* 32(18):6117–6125

Fleming SM, Lau HC. 2014. How to measure metacognition. *Frontiers in Human Neuroscience* 8:443

Fleming SM, Maniscalco B, Ko Y, Amendi N, Ro T, Lau HC. 2015. Action-specific disruption of perceptual confidence. *Psychological Science* 26(1):89–98

Fleming SM, Ryu J, Golfinos JG, Blackmon KE. 2014. Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain* 137(Pt 10):2811–2822

Fleming SM, van der Putten EJ, Daw ND. 2018. Neural mediators of changes of mind about perceptual decisions. *Nature Neuroscience* 21(4):617–624

Fleming SM, Weil RS, Nagy Z, Dolan RJ, Rees G. 2010. Relating introspective accuracy to individual differences in brain structure. *Science (New York, N.Y.)* 329(5998):1541–1543

Fu Z, Beam D, Chung JM, Reed CM, Mamelak AN, et al. 2022. The geometry of domain-general performance monitoring in the human medial frontal cortex. *Science (New York, N.Y.)* 376(6593):eabm9922

Gallagher RM, Suddendorf T, Arnold DH. 2019. Confidence as a diagnostic tool for perceptual aftereffects. *Scientific Reports* 9(1):7124

Geurts LS, Cooke JRH, van Bergen RS, Jehee JFM. 2022. Subjective confidence reflects representation of Bayesian probability in cortex. *Nature Human Behaviour* 6(2):294–305

Gherman S, Philiastides M. 2018. Human VMPFC encodes early signatures of confidence in perceptual decisions. *eLife* 7:e38293

Goupil L, Kouider S. 2016. Behavioral and Neural Indices of Metacognitive Sensitivity in Preverbal Infants. *Current Biology* 26(22):3038–3045

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

Guggenmos M. 2022. Reverse engineering of metacognition. *eLife* 11:e75420

Guggenmos M, Wilbertz G, Hebart MN, Sterzer P. 2016. Mesolimbic confidence signals guide perceptual learning in the absence of external feedback. *eLife* 5:345

Hauser TU, Allen M, Purg N, Moutoussis M, Rees G, Dolan RJ. 2017. Noradrenaline blockade specifically enhances metacognitive performance. *eLife* 6:468

Hertz U, Palminteri S, Brunetti S, Olesen C, Frith CD, Bahrami B. 2017. Neural computations underpinning the strategic management of influence in advice giving. *Nature Communications* 8(1):1–12

Heyes C, Bang D, Shea N, Frith CD, Fleming SM. 2020. Knowing Ourselves Together: The Cultural Origins of Metacognition. *Trends in Cognitive Sciences* 24(5):349–362

Hoven M, Lebreton M, Engelmann JB, Denys D, Luigjes J, van Holst RJ. 2019. Abnormalities of confidence in psychiatry: An overview and future perspectives. *Translational Psychiatry* 9(1):1–18

Hu X, Liu Z, Li T, Luo L. 2015. Influence of cue word perceptual information on metamemory accuracy in judgement of learning. *Memory* 24(3):383–398

Hu X, Zheng J, Su N, Fan T, Yang C, et al. 2021. A Bayesian inference model for metamemory. *Psychological Review* 128(5):824–855

Jiang S, Wang S, Wan X. 2022. Metacognition and mentalizing are associated with distinct neural representations of decision uncertainty. *PLoS biology* 20(5):e3001301

Kepecs A, Mainen ZF. 2012. A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society B-Biological Sciences* 367(1594):1322–1337

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

Kersten D, Mamassian P, Yuille A. 2004. Object perception as Bayesian inference. *Annual review of psychology* 55:271–304

Khalvati K, Kiani R, Rao RPN. 2021. Bayesian inference with incomplete knowledge explains perceptual confidence and its deviations from accuracy. *Nature Communications* 12(1):5704

Kiani R, Corthell L, Shadlen MN. 2014. Choice Certainty Is Informed by Both Evidence and Decision Time. *Neuron* 84(6):1329–1342

Kiani R, Shadlen MN. 2009. Representation of Confidence Associated with a Decision by Neurons in the Parietal Cortex. *Science* 324(5928):759–764

Kitayama S, Markus HR, Matsumoto H, Norasakkunkit V. 1997. Individual and collective processes in the construction of the self: self-enhancement in the united states and self-criticism in japan. *Journal of personality and social psychology* 72(6):1245

Koriat A. 1993. How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review* 100(4):609–639

Koriat A. 2012. The self-consistency model of subjective confidence. *Psychological Review* 119(1):80–113

Kwok SC, Cai Y, Buckley MJ. 2019. Mnemonic Introspection in Macaques Is Dependent on Superior Dorsolateral Prefrontal Cortex But Not Orbitofrontal Cortex. *The Journal of Neuroscience* 39(30):5922–5934

Lak A, Costa GM, Romberg E, Koulakov AA, Mainen ZF, 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

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

Lee ALF, de Gardelle V, Mamassian P. 2021. Global visual confidence. *Psychonomic Bulletin & Review* 28(4):1233–1242

Li HH, Sprague TC, Yoo AH, Ma WJ, Curtis CE. 2021. Joint representation of working memory and uncertainty in human cortex. *Neuron* 109(22):3699–3712.e6

Logan GD, Crump MJC. 2010. Cognitive illusions of authorship reveal hierarchical error detection in skilled typists. *Science* 330(6004):683–686

Ma W, Beck J, Latham P, Pouget A. 2006. Bayesian inference with probabilistic population codes. *Nature Neuroscience* 9(11):1432–1438

Mamassian P, de Gardelle V. 2022. Modeling perceptual confidence and the confidence forced-choice paradigm. *Psychological Review* 129(5):976–998

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

Maniscalco B, Odegaard B, Grimaldi P, Cho SH, Basso MA, et al. 2021. Tuned inhibition in perceptual decision-making circuits can explain seemingly suboptimal confidence behavior. *PLoS computational biology* 17(3):e1008779

Marcke HV, Denmat PL, Verguts T, Desender K. 2022. Manipulating prior beliefs causally induces under- and over-confidence, p. 10.1101/2022.03.01.482511, bioRxiv

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.e18

Mazancieux A, Fleming SM, Souchay C, Moulin CJA. 2020. Is There a G Factor for Metacognition? Correlations in Retrospective Metacognitive Sensitivity Across Tasks. *Journal of Experimental Psychology. General* 149(9):1788–1799

McCurdy LY, Maniscalco B, Metcalfe J, Liu KY, de Lange FP, Lau HC. 2013. Anatomical coupling between distinct metacognitive systems for memory and visual perception. *Journal of Neuroscience* 33(5):1897–1906

Metcalfe J, Shimamura AP, eds. 1994. Metacognition: Knowing about knowing. Metacognition: Knowing about Knowing. Cambridge, MA, US: The MIT Press

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

Middlebrooks PG, Sommer MA. 2012. Neuronal correlates of metacognition in primate frontal cortex. *Neuron* 75(3):517–530

Miller EK, Cohen JD. 2001. An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience* 24:167–202

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

Miyamoto K, Setsuie R, Osada T, Miyashita Y. 2018. Reversible Silencing of the Frontopolar Cortex Selectively Impairs Metacognitive Judgment on Non-experience in Primates. *Neuron* 97(4):980–989.e6

Miyoshi K, Lau H. 2020. A decision-congruent heuristic gives superior metacognitive sensitivity under realistic variance assumptions. *Psychological Review* 127(5):655–671

Morales J, Lau H, Fleming SM. 2018. Domain-General and Domain-Specific Patterns of Activity Supporting Metacognition in Human Prefrontal Cortex. *The Journal of Neuroscience* 38(14):3534–3546

Morrison J. 2016. Perceptual Confidence. *Analytic Philosophy* 57(1):15–48

Murphy PR, Robertson IH, Harty S, O'Connell RG. 2015. Neural evidence accumulation persists after choice to inform metacognitive judgments. *eLife* 4:3478

Nelson TO, Narens L. 1990. Metamemory: A theoretical framework and new findings. *The psychology of learning and motivation: Advances in research and theory* 26:125–173

Nicholson T, Williams DM, Lind SE, Grainger C, Carruthers P. 2021. Linking metacognition and mindreading: Evidence from autism and dual-task investigations. *Journal of Experimental Psychology. General* 150(2):206–220

Nieuwenhuis S, Ridderinkhof KR, Blom J, Band GP, Kok A. 2001. Error-related brain potentials are differentially related to awareness of response errors: Evidence from an antisaccade task. *Psychophysiology* 38(05):752–760

Norman DA, Shallice T. 1986. Attention to action: Willed and automatic control of behaviour. In *Consciousness and Self-Regulation Advances in Research and Theory*, eds. RJ Davidson, GE Schwartz, D Shapiro. wexler.free.fr, 1–18

Pannu J, Kaszniak A. 2005. Metamemory experiments in neurological populations: A review. *Neuropsychology Review* 15(3):105–130

Pereira M, Faivre N, Iturrate I, Wirthlin M, Serafini L, et al. 2020. Disentangling the origins of confidence in speeded perceptual judgments through multimodal imaging. *Proceedings of the National Academy of Sciences* 117(15):8382–8390

Pleskac TJ, Busemeyer JR. 2010. Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review* 117(3):864–901

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

Purcell BA, Kiani R. 2016. Hierarchical decision processes that operate over distinct timescales underlie choice and changes in strategy. *Proceedings of the national academy of sciences* 113(31):E4531–40

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

Rahnev D. 2021. Visual metacognition: Measures, models, and neural correlates. *The American Psychologist* 76(9):1445–1453

Rahnev D, Balsdon T, Charles L, de Gardelle V, Denison R, et al. 2022. Consensus Goals in the Field of Visual Metacognition. *Perspectives on Psychological Science* 17(6):1746–1765

Rahnev D, Denison RN. 2018. Suboptimality in perceptual decision making. *The Behavioral and Brain Sciences* 41:e223

Rahnev D, Desender K, Lee ALF, Adler WT, Aguilar-Lleyda D, et al. 2020. The Confidence Database. *Nature Human Behaviour* 4(3):317–325

Resulaj A, Kiani R, Wolpert DM, Shadlen MN. 2009. Changes of mind in decision-making. *Nature* 461(7261):263–266

Rollwage M, Dolan RJ, Fleming SM. 2018. Metacognitive Failure as a Feature of Those Holding Radical Beliefs. *Current Biology* 28(24):4014–4021.e8

Rouault M, Dayan P, Fleming SM. 2019. Forming global estimates of self-performance from local confidence. *Nature Communications* 10(1):1–11

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

Rouault M, McWilliams A, Allen MG, Fleming SM. 2018a. Human metacognition across domains: Insights from individual differences and neuroimaging. *Personality Neuroscience* 1

Rouault M, Seow T, Gillan CM, Fleming SM. 2018b. Psychiatric Symptom Dimensions Are Associated With Dissociable Shifts in Metacognition but Not Task Performance. *Biological Psychiatry* 84(6):443–451

Rouy M, de Gardelle V, Reyes G, Sackur J, Vergnaud JC, et al. 2022. Metacognitive improvement: Disentangling adaptive training from experimental confounds. *Journal of Experimental Psychology: General* 151:2083–2091

Rutishauser U, Ye S, Koroma M, Tudusciuc O, Ross IB, et al. 2015. Representation of retrieval confidence by single neurons in the human medial temporal lobe. *Nature Neuroscience* 18(7):1041–1050

Sanders JI, Hangya B, Kepecs A. 2016. Signatures of a Statistical Computation in the Human Sense of Confidence. *Neuron* 90(3):499–506

Sarafyazd M, Jazayeri M. 2019. Hierarchical reasoning by neural circuits in the frontal cortex. *Science* 364(6441):eaav8911

Schulz L, Rollwage M, Dolan RJ, Fleming SM. 2020. Dogmatism manifests in lowered information search under uncertainty. *Proceedings of the National Academy of Sciences* 117(49):31527–31534

Seow TXF, Rouault M, Gillan CM, Fleming SM. 2021. How Local and Global Metacognition Shape Mental Health. *Biological Psychiatry* 90(7):436–446

Sepulveda P, Usher M, Davies N, Benson AA, Ortoleva P, De Martino B. 2020. Visual attention modulates the integration of goal-relevant evidence and not value. *eLife* 9:e60705

Shea N, Boldt A, Bang D, Yeung N, Heyes C, Frith CD. 2014. Supra-personal cognitive control and metacognition. *Trends in Cognitive Sciences*

Shea N, Frith CD. 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

Shekhar M, Rahnev D. 2021. Sources of Metacognitive Inefficiency. *Trends in Cognitive Sciences* 25(1):12–23

Siedlecka M, Paulewicz B, Wierzchoń M. 2016. But I Was So Sure! Metacognitive Judgments Are Less Accurate Given Prospectively than Retrospectively. *Frontiers in Psychology* 7(240):218

Spence ML, Dux PE, Arnold DH. 2016. Computations underlying confidence in visual perception. *Journal of Experimental Psychology. Human Perception and Performance* 42(5):671–682

Trommershauser J, Maloney LT, Landy MS. 2008. Decision making, movement planning and statistical decision theory. *Trends in cognitive sciences* 12(8):291–297

Tsujimoto S, Genovesio A, Wise SP. 2010. Evaluating self-generated decisions in frontal pole cortex of monkeys. *Nature Neuroscience* 13(1):120–126

Urai AE, Braun A, Donner TH. 2017. Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nature Communications* 8:14637

Vaccaro AG, Fleming SM. 2018. Thinking about thinking: A coordinate-based meta-analysis of neuroimaging studies of metacognitive judgements. *Brain and Neuroscience Advances* 2:2398212818810591

van den Berg R, Anandalingam K, Zylberberg A, Kiani R, Shadlen MN, Wolpert DM. 2016a. A common mechanism underlies changes of mind about decisions and confidence. *eLife* 5:e12192

van den Berg R, Zylberberg A, Kiani R, Shadlen MN, Wolpert DM. 2016b. Confidence Is the Bridge between Multistage Decisions. *Current Biology: CB* 26(23):3157–3168

van der Plas E, Zhang S, Dong K, Bang D, Li J, et al. 2022. Identifying cultural differences in metacognition. *Journal of Experimental Psychology: General* 151(12):3268–3280

Walker EY, Pohl S, Denison RN, Barack DL, Lee J, et al. 2022. Studying the neural representations of uncertainty

Webb TW, Miyoshi K, So TY, Rajananda S, Lau H. 2022. Natural statistics support a rational account of confidence biases, p. 10.1101/2021.09.28.462081, bioRxiv

West RK, Harrison WJ, Matthews N, Mattingley JB, Sewell DK. 2022. Modality Independent or Modality Specific? Common Computations Underlie Confidence Judgements in Visual and Auditory Decisions

Wilimzig C, Tsuchiya N, Fahle M, Einhäuser W, Koch C. 2008. Spatial attention increases performance but not subjective confidence in a discrimination task. *Journal of Vision* 8(5):7.1–10

Winter CJ, Peters MAK. 2022. Variance misperception under skewed empirical noise statistics explains overconfidence in the visual periphery. *Attention, Perception & Psychophysics* 84(1):161–178

Wittmann MK, Kolling N, Faber NS, Scholl J, Nelissen N, Rushworth MFS. 2016. Self-Other Mergence in the Frontal Cortex during Cooperation and Competition. *Neuron* 91(2):482–493

Wittmann MK, Trudel N, Trier HA, Klein-Flügge MC, Sel A, et al. 2021. Causal manipulation of self-other mergence in the dorsomedial prefrontal cortex. *Neuron* 109(14):2353–2361.e11

Wokke ME, Achoui D, Cleeremans A. 2020. Action information contributes to metacognitive decision-making. *Scientific Reports* 10(1):3632

Wright ND, Edwards T, Fleming SM, Dolan RJ. 2012. Testosterone induces off-line perceptual learning. *Psychopharmacology* 224(3):451–457

Xue K, Shekhar M, Rahnev D. 2021. Examining the robustness of the relationship between metacognitive efficiency and metacognitive bias. *Consciousness and Cognition* 95:103196

Yates JF, Lee JW, Shinotsuka H, Patalano AL, Sieck WR. 1998. Cross-cultural variations in probability judgment accuracy: Beyond general knowledge overconfidence? *Organizational behavior and human decision processes* 74(2):89–117

Ye Q, Zou F, Lau H, Hu Y, Kwok SC. 2018. Causal Evidence for Mnemonic Metacognition in Human Precuneus. *Journal of Neuroscience* 38(28):6379–6387

Yeon J, Shekhar M, Rahnev D. 2020. Overlapping and unique neural circuits are activated during perceptual decision making and confidence. *Scientific Reports* 10(1):20761

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

Yu A, Dayan P. 2005. Uncertainty, neuromodulation, and attention. *Neuron* 46(4):681–692

Zheng Y, Wang D, Ye Q, Zou F, Li Y, Kwok SC. 2021. Diffusion property and functional connectivity of superior longitudinal fasciculus underpin human metacognition. *Neuropsychologia* 156:107847

Zylberberg A, Barttfeld P, Sigman M. 2012. The construction of confidence in a perceptual decision. *Frontiers in Integrative Neuroscience* 6:79

Zylberberg A, Roelfsema PR, Sigman M. 2014. Variance misperception explains illusions of confidence in simple perceptual decisions. *Consciousness and Cognition* 27:246–253