Opinion

Sources of Metacognitive Inefficiency

Medha Shekhar[1,*] and Dobromir Rahnev[1,*]

Confidence judgments are typically less informative about one's accuracy than they could be; a phenomenon we call metacognitive inefficiency. We review the existence of different sources of metacognitive inefficiency and classify them into four categories based on whether the corruption is due to: (i) systematic or nonsystematic influences, and (ii) the input to or the computation of the metacognitive system. Critically, the existence of different sources of metacognitive inefficiency provides an alternative explanation for behavioral findings typically interpreted as evidence for domain-specific (and against domain-general) metacognitive systems. We argue that, contrary to the dominant assumption in the field, metacognitive failures are not monolithic and suggest that understanding the sources of metacognitive inefficiency should be a primary goal of the science of metacognition.

Highlights

Many studies have shown that confidence ratings are less informative than they could be, a phenomenon we refer to as metacognitive inefficiency.

We review the sources of metacognitive inefficiency and classify them into four different categories, such that each source is either nonsystematic (random and nonpredictable) or systematic (predictable), and corrupts either the input to the confidence computation or the confidence computation itself.

We suggest ways to determine the relative contribution of different sources to the overall metacognitive inefficiency observed on a given task.

The existence of multiple independent sources of metacognitive inefficiency leads to a re-interpretation of studies that rely on the correlation between metacognitive scores to determine whether metacognition is domain general or domain specific.

Inefficiency in Metacognition

Humans have the **metacognitive ability** (see Glossary) to estimate the accuracy of their decisions via **confidence ratings** [1]. Metacognitive ability is reflected by the extent to which a person's confidence ratings predict their objective performance on a task [2]. However, metacognitive judgments are not always reliable indicators of objective performance. We call this phenomenon **metacognitive inefficiency**. Metacognitive inefficiency occurs when confidence judgments are less informative about the accuracy of a decision than they could be. Many studies have explored the nature of this inefficiency. For example, substantial progress has been made in revealing the neural correlates of metacognitive inefficiency [3–12] and in understanding whether there are stable individual differences in metacognitive inefficiency [8,13–17]. Nevertheless, little attention has been paid to whether these efforts identify the same sources of metacognitive inefficiency or how the different sources should be classified.

Here, we explore the different sources of metacognitive inefficiency and classify them into four different categories based on the distinctions between systematic versus nonsystematic corruption, as well as failures of input versus computation. We argue that different tasks are likely dominated by different sources of inefficiency and that ignoring this fact can lead to incorrect conclusions. While this opinion article focuses on understanding the (in)efficiency in the monitoring function of metacognition [18,19], we note that metacognition has additional functions such as control and social communication [18–22] and, in some cases, inefficiency in monitoring may arise from the efficient function of these other functions of metacognition or even from processes that are well adapted to the real world but not to the laboratory. Additionally, we focus on metacognition in perceptual decision making but our conclusions apply equally well to metacognition in other domains such as memory, social, and value-based decisions.

Evidence for Metacognitive Inefficiency

Before we examine the different sources of metacognitive inefficiency, we briefly review the evidence for metacognitive inefficiency. One can identify both **model-based** and **model-free evidence** for metacognitive inefficiency; that is, evidence that is or is not derived from a computational model. Model-based evidence stems primarily from two-choice tasks that are well

[1] School of Psychology, Georgia Institute of Technology, Atlanta, GA, USA

*Correspondence:
medha@gatech.edu (M. Shekhar) and
rahnev@psych.gatech.edu (D. Rahnev).

https://doi.org/10.1016/j.tics.2020.10.007

described by **signal detection theory (SDT)**. SDT can provide a theoretical ideal for the maximum informativeness of confidence judgments for these types of tasks in the form of stimulus sensitivity, d′. Using SDT assumptions, one can derive a measure of the informativeness of confidence ratings, called meta-d′, that is in the same units as the stimulus sensitivity, d′ [23]. For observers with ideal metacognitive efficiency, meta-d′ should equal d′, implying that confidence ratings are as informative as the primary decision. Therefore, observations of meta-d′ < d′ are typically taken as evidence for the presence of metacognitive inefficiency. Such observations are common [3,7,24–31], suggesting widespread metacognitive inefficiency. Critically, examining the data from individual subjects demonstrates that within the same task, some people are able to achieve metacognitive efficiency very close to meta-d′ = d′, whereas many others fall far below this ideal [3,32,33]. Such findings suggest that the ideal prescribed by SDT is indeed attainable, and therefore values of meta-d′ substantially lower than d′ are a sign of metacognitive inefficiency. (It should be noted that values of meta-d′ > d′ have also been observed but the reason behind them remain ill understood [34] and therefore are not further explored here.)

Beyond model-based demonstrations, there are many findings of confidence–accuracy dissociations that provide model-free evidence for metacognitive inefficiency. Indeed, a number of papers have demonstrated the existence of experimental conditions matched on accuracy but with different levels of confidence [28,35–41]. If metacognition were perfectly efficient, higher confidence should always be associated with higher accuracy and therefore such confidence–accuracy dissociations would not exist. Thus, the presence of metacognitive inefficiency is already well established with both model-based and model-free methods. What is less clear, however, is where this inefficiency comes from.

Previously Proposed Sources of Metacognitive Inefficiency

Given the importance of understanding and improving the quality of confidence judgments, it is perhaps surprising that little attention has been paid to distinguishing and systematizing the sources of metacognitive inefficiency. What is worse, the lack of attention to this issue has resulted in an implicit assumption that metacognitive inefficiency is monolithic such that all failures of metacognition are effectively the same. Here, we first review the previously proposed sources of metacognitive inefficiency as they have been identified in the literature and then develop a new system for classifying these sources.

Random Noise That Selectively Affects Confidence Judgments

Metacognitive inefficiency can occur when confidence is influenced by random noise that does not affect the perceptual decision. These noise sources can affect either the signal used for confidence or the confidence computation itself.

Noise in the Signal for Confidence

Perhaps the most widely proposed source of metacognitive inefficiency is the existence of random noise in the information used for confidence computation [5,7,25,26,29,38,42–44]. For example, decreases in metacognitive efficiency due to transcranial magnetic stimulation have been theorized to arise from increased noise in the signal used to determine confidence [7,38,42]. Similarly, many models have accounted for metacognitive inefficiency by postulating signal decay [29,44–46] or random noise [26,29,33,43] in the signal used to generate metacognitive judgments.

Noise in the Confidence Computations

A related source of metacognitive inefficiency is random noise in the confidence computations. One example is **criterion jitter** [47]: the inability of observers to maintain stable confidence

Glossary

Confidence ratings: subjective ratings indicating how certain subjects are in their response. Confidence ratings are often assumed to indicate the probability of being correct but recent works shows that this may not always be the case.

Criterion jitter: tendency for criteria that should remain stable over the course of an experiment to shift from trial to trial. Jitter in confidence criteria leads to lower informativeness of the confidence ratings and is a type of metacognitive noise.

Domain general versus domain specific: the idea that the same system (domain general) or different systems (domain specific) are responsible for the performance across two or more tasks. Whether different metacognitive tasks are domain general or domain specific has been typically assessed by correlating the metacognitive scores across subjects on these tasks.

Metacognitive ability: ability of the subject to use metacognitive ratings of confidence to differentiate between correct and incorrect responses. Higher metacognitive ability means that one's confidence ratings are more informative of their own accuracy.

Metacognitive inefficiency: phenomenon whereby metacognitive ratings of confidence are less informative about the accuracy of a decision that they could be.

Metacognitive noise: type of noise that affects confidence ratings but not the primary decision. Metacognitive noise is a feature of many recent models of confidence. Its existence leads to counterintuitive predictions that have recently been empirically validated [26].

Model-based evidence: evidence for a phenomenon that comes from testing the assumptions of a computational model. For example, model-based evidence for metacognitive inefficiency that comes from findings of meta-d′ < d′, where both meta-d′ and d′ are measures derived from SDT, which predicts that meta-d′ = d′ for ideal observers.

Model-free evidence: evidence for a phenomenon that does not depend on the assumptions of specific computational models. For example, task manipulations that affect confidence but not accuracy (or vice versa) constitute model-free evidence for metacognitive inefficiency because for ideal observers, a given confidence level must always map to the same accuracy criteria over trials [48]. This source of metacognitive inefficiency can be very difficult to distinguish from a noisy signal and sometimes the two sources of noise are mathematically equivalent [33]. Several cognitive factors like fatigue and multitasking may induce noisy confidence computations (or noisy signal for confidence) but the evidence for this is currently mixed [27,49,50].

Nonrandom, Systematic Factors That Differentially Affect Confidence and Accuracy

Metacognitive inefficiency can arise from systematic sources that differentially affect confidence and accuracy, thus decreasing the correspondence between the two judgments. Below, we discuss three possible scenarios by which such factors can cause inefficiency.

Factors That Affect Confidence Judgments but not Accuracy

Many examples of confidence judgments being influenced by factors unrelated to accuracy exist. For instance, the confidence judgment on the current trial is affected by the confidence on previous trials [48,51,52]. This confidence leak phenomenon occurs even when the previous confidence was given for a different task [51] or when confidence judgments were not elicited explicitly [53]. Such previous confidence judgments have virtually no influence on the accuracy of the current trial and thus should have been ignored.

Additionally, confidence has been found to be modulated by several other nonperceptual factors including physiological variables such as one's level of arousal [54,55]; highly accessible perceptual cues such as the font size in which to-be-remembered words are printed [56]; the contrast in which to-be-remembered pictures are presented [57]; and the stimulus uncertainty in a dimension irrelevant to the current decision [40,58–63]. In all of these cases, the critical factor affected the confidence ratings but not the task accuracy.

Other examples of confidence incorporating factors not predictive of accuracy come from findings that metacognitive judgments are influenced by how we act, independent of the content of the initial perceptual decision. For example, outputs of motor areas have been shown to inform confidence judgments. Indeed, transcranial magnetic stimulation delivered to the premotor cortex decreased both confidence and metacognitive efficiency independent of task performance [25]. Similarly, subthreshold motor activations before a planned response, which are thought to proxy corrected motor plans, correlated positively with confidence but not accuracy [64]. Further studies have shown that confidence is affected by manipulations of movement-related parameters such as increasing movement speed [65] and extending response times without affecting accuracy [66]. In all of these cases, manipulations related to how an action is executed or internally represented changed the confidence ratings without affecting the accuracy on the task, thus resulting in metacognitive inefficiency.

Factors That Affect Accuracy but not Confidence Judgments

In addition to confidence being influenced by factors unrelated to accuracy, another source of metacognitive inefficiency is confidence not being influenced enough by factors that do affect accuracy. By ignoring, or at least not fully incorporating such factors, confidence ratings become less informative and metacognitive efficiency decreases. Perhaps the best-known example of this phenomenon is the positive evidence bias [30,41,67–69]. This bias consists in the tendency for confidence ratings to disproportionately weigh the evidence in favor of a decision while ignoring evidence that is incongruent with that decision. For example, one study [68] showed that although perceptual choices are characterized by symmetric contributions of evidence for and against a choice, confidence is strongly modulated by evidence for the chosen response and completely insensitive to evidence against the chosen response. A related phenomenon is the finding that, in some situations, confidence is insensitive to signal variability even though this

level irrespective of the assumed computational model.

Signal detection theory (SDT): theory of perceptual decision making used to model choice behavior (often in two-choice tasks) relating choice behavior to the way sensory information is represented internally.

variability affects task accuracy [70]. Finally, confidence is unaffected by masked stimuli even though they affect performance on the task [28].

Incorrect Weighting of Sensory Signals for Confidence

Finally, it is also possible that confidence judgments weigh inappropriately different parts of the sensory signal itself. Indeed, in some situations, the signal on which the primary decision and confidence are based may itself consist of many different individual signals. This is the case when multiple stimuli are being judged but is also true for individual stimuli under certain classes of models such as models based on sequential sampling [71]. In such cases, it is possible that confidence is based only on a subset of all of the available signals that affected the accuracy on the task. For example, confidence ratings may consider only the last part of the evidence accumulation process or only on the evidence of some accumulators but not others [72,73].

Confidence Computing the Wrong Quantity

To be maximally predictive of accuracy, confidence computations should be such that the final confidence rating is given by placing criteria directly on the axis of the probability of being correct, although the criteria themselves may be biased (resulting in over- or underconfidence). Following previous literature [63,74–76], we refer to such confidence computation as ‘reflecting’ probability correct. Several studies using two-choice tasks have found that, for the majority of subjects, confidence does not reflect probability correct but rather reflects heuristic decision rules that approximate but do not compute the posterior probability of being correct [75,77]. Similarly, a recent paper examined confidence in three-choice tasks and found that subjects do not base their confidence on the probability of being correct in such tasks; instead, confidence appears to reflect the difference in posterior probability between the most likely and the second most likely option [74].

Methodological Issues That Masquerade as Metacognitive Inefficiency

Finally, in addition to the genuine sources of metacognitive inefficiency highlighted above, there are certain methodological factors that, if ignored, can masquerade as metacognitive inefficiency. For example, subjects may strategically shift their confidence criteria as they are learning a task (e.g., if they were instructed to use the whole confidence scale and they realize midway through the experiment that they are not) or adapting to external conditions such as communicating with a partner [20,22]. Additionally, some sources of inefficiency may only be problematic in a laboratory setting but still constitute appropriate heuristics outside the laboratory (Box 1). Conversely, many factors – such as inattention, lack of motivation, unclear instructions, or unintuitive confidence scales – are sometimes informally considered as sources of metacognitive inefficiency but in fact cannot directly corrupt the confidence ratings. Instead, to the extent to which these factors contribute to metacognitive inefficiency, they necessarily do so via the same computational mechanisms highlighted above (e.g., lack of motivation might increase criterion jitter or the noise in the sensory signal for confidence).

Categorizing the Sources of Metacognitive Inefficiency

As the brief overview above shows, a number of sources of metacognitive inefficiency have already been identified and there are likely even more sources yet to be discovered. The large number of individual sources described in the literature makes it necessary that we identify critical dimensions that can help us better understand the nature of these sources of metacognitive inefficiency. We propose that existing sources of metacognitive inefficiency could be understood based on two key considerations: (i) does the corruption arise from systematic (predictable) or nonsystematic (random) causes; and (ii) is the corruption due to the input to the confidence

Box 1. The Ultimate Reasons for Metacognitive Inefficiency

Given the importance of confidence in the real world [19,39,94–97], it is natural to ask why metacognitive inefficiency was not eliminated by evolution.

If they exist, nonsystematic sources of metacognitive inefficiency are likely caused by inherent processing or resource limitations [27,50]. It is possible that random input noise is due to inevitable signal corruption associated with passing information from decision circuits to circuits involved in confidence computations. Similarly, it could be that neural circuits cannot perform noiseless computations and phenomena like criterion jitter are inevitable. These two categories of metacognitive inefficiency can thus be expected to be similar across different tasks and to have a detrimental effect in both the real world and the laboratory.

However, the ultimate cause of the systematic sources of metacognitive inefficiency is likely different. Of course, it is possible that some of them are also due to processing or resource limitations, especially when computing the probability of being correct is complex or resource demanding [75,77]. However, in many cases, it is more likely that the ultimate reason for these sources of metacognitive inefficiency is that confidence is based on mechanisms that are well adapted to the real world but not to the laboratory. For instance, unlike in the laboratory, decisions in the real world are rarely between two discrete alternatives [74,98] and estimating the evidence associated with each of many possible alternatives is inefficient and often impossible. Therefore, in such situations, confidence may be based exclusively on the evidence for the most likely option, which could explain findings of positive evidence bias [30]. Similarly, in the real world, there is a high degree of temporal continuity between events that can be exploited by our decision-making systems, which could explain the confidence leak phenomenon [51]. Finally, it may be beneficial if confidence in the real world reflects other quantities and not just the accuracy of the primary decision [30,67,74].

Therefore, the metacognitive inefficiency observed in the laboratory may be due both to inherent limitations of the system and to the fact that confidence has evolved to be used in different situations than the ones commonly used in the laboratory. It remains an open question which set of factors ultimately has a higher contribution to metacognitive inefficiency.

computation or due to the confidence computation itself. Taken together, these two dimensions create four categories of metacognitive inefficiency (Figure 1, Key Figure).

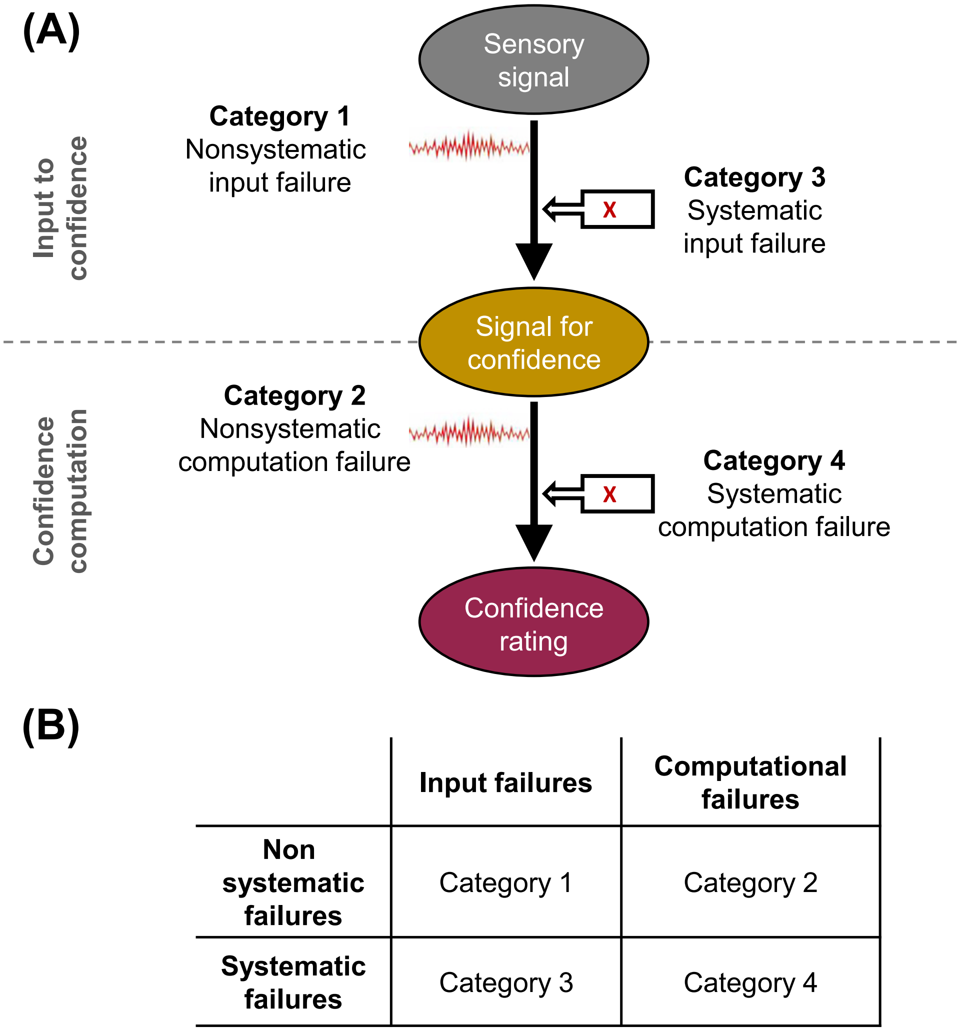
Nonsystematic versus Systematic Sources of Inefficiency

In the first categorization, we can distinguish between nonsystematic (Categories 1 and 2) and systematic (Categories 3 and 4) sources of metacognitive inefficiency. The critical difference is that the influence of the systematic sources can be predicted on a trial-by-trial basis, whereas the influence of the nonsystematic sources cannot. For example, the presence of high positive evidence or high confidence on the previous trial can be used to predict that confidence on the current trial would be relatively high, which makes these phenomena systematic sources of metacognitive inefficiency. By contrast, the presence of random noise can tell us to expect a noisy confidence judgment but does not allow us to predict if the confidence rating would be low or high, which makes random noise a nonsystematic source of metacognitive inefficiency. Among the sources reviewed above, noise in the signal for confidence (equivalent to Category 1 here) and noise in the confidence computation (equivalent to Category 2 here) are nonsystematic, whereas all of the other sources (which fall under Categories 3 and 4 here) are systematic.

At present, there is a strong bias towards viewing metacognitive inefficiency as primarily due to nonsystematic corruption. Indeed, as already mentioned, random noise in the signal for confidence (Category 1) is perhaps the most ubiquitously assumed source of metacognitive inefficiency and appears in many models of confidence [5,7,25,26,29,38,42–44]. However, there is yet no empirical evidence that any of the observed metacognitive inefficiency stems from random, nonpredictable sources. Instead, it might be that metacognitive inefficiency is only caused by systematic sources. However, if these sources are not measured or modeled, then the only way of capturing their influence is by postulating a nonsystematic source of noise. For example, confidence may systematically change based on the confidence on the previous

Key Figure

Categorizing the Sources of Metacognitive Inefficiency



	Input failures	Computational failures
Non systematic failures	Category 1	Category 2
Systematic failures	Category 3	Category 4

Trends in Cognitive Sciences

Figure 1. (A) Metacognitive inefficiency may arise from either systematic or nonsystematic sources, as well as from failures in either input or computation. (B) Combining these two dimensions results in four categories of metacognitive inefficiency. Nonsystematic sources of metacognitive inefficiency (Categories 1 and 2) lead to random perturbations that cannot be used to predict confidence on a trial-by-trial basis, whereas systematic sources of metacognitive inefficiency (Categories 3 and 4) lead to predictable perturbations that can be used to predict confidence on a trial-by-trial basis. Input failures (Categories 1 and 3) affect the signal for confidence such that confidence ratings are bound to be less informative regardless of the computation, whereas computational failures (Categories 2 and 4) are due to the metacognitive system arriving at less informative confidence ratings despite working with signals that allow for more informative confidence ratings to be made. Several categories of failures can coexist within a single task. Metacognitive inefficiency is currently most often viewed as input failure (Categories 1 and 3) but in virtually all cases the empirical findings could also be explained as computational failure instead (Categories 2 and 4). Similarly, many models include nonsystematic sources of noise (Categories 1 and 2) but it is currently unclear whether confidence judgments are corrupted by truly random noise or by unmodeled systematic sources of metacognitive inefficiency (Categories 3 and 4).

trial or the variance of the signal on the current trial, but, if these quantities are not modeled, then it will appear as if confidence ratings are corrupted in a random fashion.

Therefore, at present, the inclusion of nonsystematic sources of metacognitive inefficiency in models of metacognition should not be taken literally as a substantive claim that the actual source of metacognitive inefficiency is truly random and nonpredictable. Stated differently, studies that model metacognitive inefficiency as stemming from random noise in the sensory signal [7,26,29,33,78] ought not be interpreted as evidence that this type of corruption (Category 1) is the true cause of the observed inefficiency. Ultimately, determining the role of nonsystematic sources of inefficiency would require measuring and quantifying all systematic sources of corruption; the inefficiency that is left unexplained after that can more confidently be attributed to random, nonsystematic sources.

Input versus Computational Failures

In the second categorization, we can distinguish between sources of metacognitive inefficiency related to the input to the confidence computation (Categories 1 and 3) versus the computation itself (Categories 2 and 4; Figure 1). Input failures cast metacognitive inefficiency as due to the metacognitive system operating with noisy, corrupted, or incomplete signals. If the system responsible for generating confidence ratings does not have access to the same type or quality of information as the system making the primary decision, then this would be an input failure. However, computational failures cast metacognitive inefficiency as due to computational deficiencies unrelated to the incoming signal. If the metacognitive system does have access to the same information as the system making the primary decision, but still generates confidence ratings that are not as predictive of accuracy as they could be, then this would be a computational failure.

Similar to the bias towards nonsystematic sources of metacognitive inefficiency, there appears to be a bias towards casting metacognitive inefficiency as stemming from input failures. Indeed, the majority of the systematic sources of metacognitive inefficiency reviewed above are typically seen as input failures (Category 3) but for virtually all of these sources, the corruption could easily be due to a computational failure instead (Category 4). For example, positive evidence bias, which results from the neglect of decision-incongruent information, could occur either because metacognition does not have access to decision-incongruent evidence (Category 3 failure) or because it chooses to exclude or underweight this information in its computation (Category 4 failure). Analogous to the systematic sources of corruption, it is also more common to model nonsystematic sources of metacognitive inefficiency as noise in the signal rather than noise in the confidence criteria, even though these sources of corruption are often mathematically equivalent [33]. Therefore, for both systematic and nonsystematic sources of metacognitive inefficiency, the decision on whether to cast them as input or computational failures largely depends on the bias of the individual modeler rather than direct empirical evidence. In fact, currently we can only unambiguously distinguish between input and computational failures in the cases described above as confidence computing the wrong quantity. Those sources of metacognitive inefficiency are clearly due to computational (Category 4) rather than input (Category 3) failures. Thus, despite the bias towards casting metacognitive inefficiency as stemming from input failures, we currently only have strong evidence for cases of metacognitive inefficiency stemming from computational failures.

Ultimately, adequately adjudicating between input and computation failures is likely to require the development of new neural or behavioral methods where each type of failure can be identified or manipulated. Indeed, parallel efforts within perceptual decision making have been successful in disentangling the sensory and decisional sources of corruption in perceptual decisions [79].

Determining the Importance of Each Source of Metacognitive Inefficiency

Many modeling papers subsume all sources of metacognitive inefficiency under the umbrella term of **metacognitive noise**, which is random, nonsystematic noise in the confidence ratings that is not present in the perceptual decision [7,26,29,33,43,48]. Unless different systematic sources of metacognitive inefficiency are measured and modeled, lumping all sources into a single nonsystematic noise term is perhaps the best we could do. However, many systematic sources of metacognitive inefficiency have been identified and they can indeed be measured and modeled [30,45,46,51,67,74,75]. This gives us an opportunity to determine the extent to which each source of metacognitive inefficiency contributes to the overall inefficiency of the confidence ratings.

How can this be done in practice? A promising approach is to include multiple sources in a single model and then compare their relative contributions. Perhaps the simplest way of implementing this approach is to build a model that includes both a specific systematic source of metacognitive inefficiency and a catch-all, nonsystematic noise term. The relative contributions of each source can then be assessed based on the best fitting parameters of the model. For example, one could examine the strength of metacognitive inefficiency if one source of inefficiency is removed, and thus compare the relative contribution of each. In fact, some papers already include models with multiple sources of metacognitive inefficiency [45,74], and in such cases it should already be possible to compare the influence of the different sources of inefficiency. In other cases, models that only account for a specific systematic source of inefficiency [30,51] can be augmented with relative ease by adding an additional, nonsystematic source of noise. In addition, future studies can combine two or more of systematic sources of inefficiency and assess their relative contribution. A specific example of this approach is presented in Box 2. Making such complex models identifiable

Box 2. Determining the Influence of Different Sources of Metacognitive Inefficiency

There are many sources of metacognitive inefficiency, but it is unlikely that they are all equally important. Here we outline how one could begin to determine the strength of the influence of different sources. Specifically, we model the influence of (i) the confidence on the previous trial; (ii) the arousal on the current trial; and (iii) a catch-all, nonsystematic source of noise that captures all remaining sources of metacognitive inefficiency.

Consider a two-choice task where the stimulus s can come from two categories $s = (-1, 1)$. The signal for the decision, r_{dec}, is corrupted by Gaussian sensory noise but is not influenced by any of the three factors above, such that $r_{dec} = N(s * \mu, \sigma_s^2)$, where μ is the signal strength and σ_s^2 is the variance of the sensory noise. The decision d can be made such that $d = \begin{cases} -1, r_{dec} \leq 0 \\ 1, r_{dec} > 0 \end{cases}$. The confidence variable, r_{conf}, can then be modeled as a function of the decision, d, the decision variable, r_{dec}, and the three sources of inefficiency above. There are many possible models for the confidence variable but here we will consider an extremely simple model for illustration.

Let $conf_{prev}$ be the confidence on the previous trial, the parameter a be the level of arousal measured on the current trial, and $\epsilon = N(0, \sigma_{meta}^2)$ be a nonsystematic source of noise with zero mean and variance of σ_{meta}^2. One possible model of confidence generation is based on the following formula for the confidence variable:

$$r_{conf} = f\left(d, r_{dec}, conf_{prev}, a, \epsilon\right) = d * r_{dec} + w_1 * conf_{prev} + w_2 * a + \epsilon$$

where w_1 and w_2 are weights which have to be estimated. This model thus has three free parameters – w_1, w_2, and σ_{meta}^2 – corresponding to the strength of influence of each factor. Additional free parameters would be needed to transform r_{conf} into discrete confidence ratings. Note that this model is silent on whether the sources of metacognitive inefficiency stem from input or computational failures as both interpretations are viable.

Assuming that this model fits the data better than competing models, we can then assess the relative contribution of each source of metacognitive inefficiency. One way of doing so could be to simulate the observer's behavior by removing the influence of each of these three sources in turns from the formula for r_{conf} and determining the resulting improvement in metacognitive efficiency (meta-d/d'). This approach would tell us how much better one's metacognitive ability could have been had a specific source of metacognitive inefficiency been avoided.

requires experiments with more complex designs – such as having tasks with multiple alternatives, manipulating stimulus contrast, or measuring arousal – but many such datasets are already freely available [80]. Therefore, immediate progress on distinguishing the contributions of different sources of metacognitive inefficiency is within reach.

Implications for Existing Research

Appreciating that metacognitive inefficiency is not a monolithic phenomenon has strong implications for our understanding of domain generality of metacognition [3,6,8,13,14,16,17,32,81–85] and other cognitive skills such as intelligence, learning, and creativity [86,87]. A skill is **domain general** when the same system, consisting of a single set of neural and cognitive mechanisms, is responsible for its execution across multiple tasks. By contrast, a skill is **domain specific** if different systems are responsible for its execution in different tasks. Determining whether a process is domain general or domain specific thus has strong implications about the organization of cognition. Critically, in both metacognition and other fields, the existence of domain generality or specificity has been assessed based on intertask correlations in performance with high correlations being a sign of domain generality and low correlations being a sign of domain specificity.

However, this logic implicitly assumes a single source of corruption for all tasks. Appreciating the possibility of multiple sources of inefficiency instead offers an alternative explanation for such findings. Specifically, it could be that there is a single system for computing confidence across all tasks (that is, metacognition is fully domain general) and the correlations in metacognitive accuracy between tasks are driven by the overlap in the sources of inefficiency for the different tasks (Figure 2). In this interpretation, high intertask correlations imply that the same source of metacognitive inefficiency dominates both tasks, whereas low correlations suggest that the two tasks are dominated by different, uncorrelated sources of inefficiency. This alternative interpretation is supported by a recent study which found that using a particular task type increased the intertask correlation in metacognitive efficiency across the domains of perception and memory [82], consistent with the idea that different domains are dominated by different sources of noise but using the same task type allows similar sources of metacognitive inefficiency to prevail in both tasks. It should also be noted that a lack of correlation may also arise due to insufficient statistical power with larger studies being more likely to find significant correlations [13,82,84].

The existence of multiple sources of metacognitive inefficiency also has implications for the neural correlates of metacognitive inefficiency [3–5,51]. Previous studies have measured the correlation between metacognitive ability and various brain measures. Although this is an important first step towards understanding the neural basis of metacognitive inefficiency, the presence of an omnibus correlation does not reveal which sources of metacognitive inefficiency are mediated by the anatomical regions uncovered by such correlations. Identifying the different sources of metacognitive inefficiency in a given task can therefore allow a much more precise interpretation of the role of different brain regions, and potentially uncover additional anatomical correlates associated with sources of noise that were not dominant in previously used tasks.

Concluding Remarks

Metacognition is inefficient but, unlike dominant assumptions in the field, this inefficiency is not monolithic. Here, we review the different sources of metacognitive inefficiency and identify four different categories based on the distinction of systematic versus nonsystematic and input versus computation failures. This categorization is general and should apply to many other areas of cognition. We argue that to understand the nature of metacognitive judgments, it is critical that

Outstanding Questions

What other sources of metacognitive inefficiency exist beyond the ones discussed in this opinion article?

What are the relative contributions of each of the sources of metacognitive inefficiency? Are some sources more important than others?

How can we determine whether random, nonsystematic sources of metacognitive inefficiency truly exist or if they are just mathematical tools used to help our models fit the data better?

How can we find out whether a given source of metacognitive inefficiency affects the input to the confidence computation or the computation itself? Which of these two types of failure (input versus computation) has greater influence on the metacognitive inefficiency observed empirically?

What additional dimensions (beyond systematic/nonsystematic and input/computation) exist on which the sources of metacognitive inefficiency could be usefully classified?

Are different tasks dominated by different sources of metacognitive inefficiency? If so, how can we know in advance which task is likely to be dominated by which source of metacognitive inefficiency?

In tasks for which the corresponding metacognitive scores are uncorrelated, how can we determine if the lack of correlation is due to the presence of domain-specific mechanisms or different sources of noise that dominate the different tasks?

Do the brain areas previously shown to correlate with metacognitive ability mediate all sources of metacognitive inefficiency or only a subset of these sources?

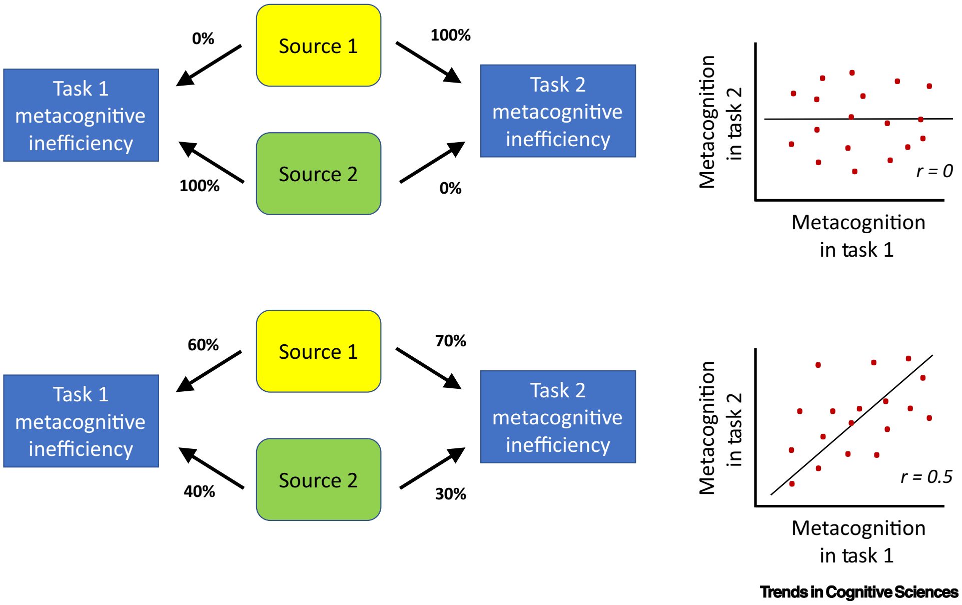

0%
Source 1
100%
Task 1 metacognitive inefficiency
Task 2 metacognitive inefficiency
100%
Source 2
0%
Metacognition in task 2
r = 0
Metacognition in task 1
60%
Source 1
70%
Task 1 metacognitive inefficiency
Task 2 metacognitive inefficiency
40%
Source 2
30%
Metacognition in task 2
r = 0.5
Metacognition in task 1
Trends in Cognitive Sciences

Figure 2. Interpreting Metacognitive Accuracy Correlations between Different Tasks. It is typically assumed that the correlation between the metacognitive scores on two different tasks can be used to infer whether metacognition is domain general (claimed in cases of positive correlation) or domain specific (claimed in cases of no correlation) [3,6,8,13,14,16,17,32,81–85]. However, a markedly different interpretation of such findings is possible. According to this interpretation, metacognition is *a priori* assumed to be domain general. A lack of correlation between two different tasks is then taken as evidence that the two tasks are dominated by different sources of metacognitive inefficiency. For example, the upper panel depicts a situation where two different sources of noise completely determine the extent of metacognitive inefficiency in Tasks 1 and 2, and therefore the correlation in the metacognitive score between these tasks is zero. By contrast, if the contribution of each source is comparable (bottom panel), then a positive correlation is observed. Thus, in this alternative interpretation, metacognition is simply assumed to be domain general and the strength of correlation between the metacognitive accuracy on two different tasks is taken as evidence regarding whether different sources of metacognitive noise dominate the two tasks.

future studies go beyond simply determining the existence of inefficiency and attempt to pinpoint its exact sources. Accomplishing this goal will likely necessitate the use of more complex tasks and models. This effort is likely to lead to surprises. We may discover that the main sources of metacognitive inefficiency are not the ones we thought (see Outstanding Questions). Identifying the exact sources of metacognitive inefficiency may also necessitate that we revise our ideas about whether metacognition is a fully domain-general process or whether some subdomains have their own dedicated metacognitive systems. Finally, a mechanistic understanding of the exact sources of metacognitive inefficiency will result in greater insight into neuropsychiatric disorders characterized by disruptions of metacognition [88–92] and potentially allow us to create targeted interventions to improve the quality of confidence ratings, especially in high-stakes situations such as eyewitness testimony [93].

Acknowledgments

This work was supported by the National Institute of Mental Health (award number: R01MH119189) and the Office of Naval Research (award number: N00014-20-1-2622).

References

1. Metcalfe, J., Shimamura, A.P., (eds) (1994) *Metacognition: Knowing about Knowing*, MIT Press
2. Baranski, J.V. and Petrusic, W.M. (1994) The calibration and resolution of confidence in perceptual judgments. *Percept. Psychophys.* 55, 412–428
3. McCurdy, L.Y. *et al.* (2013) Anatomical coupling between distinct metacognitive systems for memory and visual perception. *J. Neurosci.* 33, 1897–1906
4. Fleming, S.M. *et al.* (2010) Relating introspective accuracy to individual differences in brain structure. *Science* 329, 1541–1543

5. Fleming, S.M. *et al.* (2012) Prefrontal contributions to metacognition in perceptual decision making. *J. Neurosci.* 32, 6117–6125
6. Baird, B. *et al.* (2013) Medial and lateral networks in anterior prefrontal cortex support metacognitive ability for memory and perception. *J. Neurosci.* 33, 16657–16665
7. Shekhar, M. and Rahnev, D. (2018) Distinguishing the roles of dorsolateral and anterior PFC in visual metacognition. *J. Neurosci.* 38, 5078–5087
8. Morales, J. *et al.* (2018) Domain-general and domain-specific patterns of activity supporting metacognition in human prefrontal cortex. *J. Neurosci.* 38, 3534–3546
9. Wokke, M.E. *et al.* (2016) Sure I'm sure: prefrontal oscillations support metacognitive monitoring of decision making. *J. Neurosci.* 37, 781–789
10. Bang, D. and Fleming, S.M. (2018) Distinct encoding of decision confidence in human medial prefrontal cortex. *Proc. Natl. Acad. Sci. U. S. A.* 115, 6082–6087
11. Bor, D. *et al.* (2017) Theta-burst transcranial magnetic stimulation to the prefrontal or parietal cortex does not impair metacognitive visual awareness. *PLoS One* 12, e0171793
12. Peters, M.A.K. *et al.* (2017) Transcranial magnetic stimulation to visual cortex induces suboptimal introspection. *Cortex* 93, 119–132
13. Faivre, N. *et al.* (2018) Behavioral, modeling, and electrophysiological evidence for supramodality in human metacognition. *J. Neurosci.* 38, 263–277
14. Carpenter, J. *et al.* (2019) Domain-general enhancements of metacognitive ability through adaptive training. *J. Exp. Psychol. Gen.* 148, 51–64
15. Ais, J. *et al.* (2016) Individual consistency in the accuracy and distribution of confidence judgments. *Cognition* 146, 377–386
16. De Gardelle, V. *et al.* (2016) Confidence as a common currency between vision and audition. *PLoS One* 11, e0147901
17. Samaha, J. and Postle, B.R. (2017) Correlated individual differences suggest a common mechanism underlying metacognition in visual perception and visual short-term memory. *Proc. Biol. Sci.* 284, 20172035
18. Shimamura, A.P. (2008) A neurocognitive approach to metacognitive monitoring and control. In *Handbook of Memory and Metacognition* (Dunlosky, J. and Bjork, R.A., eds), pp. 373–390, Erlbaum Publishers
19. Nelson, T.O. (1990) Metamemory: a theoretical framework and new findings. *Psychol. Learn. Motiv. Adv. Res. Theory* 26, 125–173
20. Shea, N. *et al.* (2014) Supra-personal cognitive control and metacognition. *Trends Cogn. Sci.* 18, 186–193
21. Frith, C.D. (2012) The role of metacognition in human social interactions. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 367, 2213–2223
22. Bang, D. *et al.* (2017) Confidence matching in group decision-making. *Nat. Hum. Behav.* 1, 1–7
23. Maniscalco, B. and Lau, H. (2012) A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Conscious. Cogn.* 21, 422–430
24. Fleming, S.M. *et al.* (2014) Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain* 137, 2811–2822
25. Fleming, S.M. *et al.* (2015) Action-specific disruption of perceptual confidence. *Psychol. Sci.* 26, 89–98
26. Bang, J.W. *et al.* (2019) Sensory noise increases metacognitive efficiency. *J. Exp. Psychol. Gen.* 148, 437–452
27. Maniscalco, B. and Lau, H. (2015) Manipulation of working memory contents selectively impairs metacognitive sensitivity in a concurrent visual discrimination task. *Neurosci. Conscious.* 2015, niv002
28. Vlassova, A. *et al.* (2014) Unconscious information changes decision accuracy but not confidence. *Proc. Natl. Acad. Sci. U. S. A.* 111, 16214–16218
29. Maniscalco, B. and Lau, H. (2016) The signal processing architecture underlying subjective reports of sensory awareness. *Neurosci. Conscious.* 2016, niw002
30. Maniscalco, B. *et al.* (2016) Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity. *Attention Perception Psychophys.* 78, 923–937
31. Sherman, M.T. *et al.* (2015) Prior expectations facilitate metacognition for perceptual decision. *Conscious. Cogn.* 35, 53–65
32. Song, C. *et al.* (2011) Relating inter-individual differences in metacognitive performance on different perceptual tasks. *Conscious. Cogn.* 20, 1787–1792
33. Shekhar, M. and Rahnev, D. (2020) The nature of metacognitive inefficiency in perceptual decision making. *Psychol. Rev.* Published online July 16, 2020. https://doi.org/10.1037/rev0000249
34. Rahnev, D. and Fleming, S.M. (2019) How experimental procedures influence estimates of metacognitive ability. *Neurosci. Conscious.* 2019, niz009
35. Lau, H.C. and Passingham, R.E. (2006) Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proc. Natl. Acad. Sci. U. S. A.* 103, 18763–18768
36. Wilimzig, C. *et al.* (2008) Spatial attention increases performance but not subjective confidence in a discrimination task. *J. Vis.* 8, 7
37. Rahnev, D. *et al.* (2012) Direct injection of noise to the visual cortex decreases accuracy but increases decision confidence. *J. Neurophysiol.* 107, 1556–1563
38. Rahnev, D. *et al.* (2016) Causal evidence for frontal cortex organization for perceptual decision making. *Proc. Natl. Acad. Sci.* 113, 201522551
39. Desender, K. *et al.* (2018) Subjective confidence predicts information seeking in decision making. *Psychol. Sci.* 29, 761–778
40. Zylberberg, A. *et al.* (2016) The influence of evidence volatility on choice, reaction time and confidence in a perceptual decision. *Elife* 5, e17688
41. Samaha, J. *et al.* (2016) Dissociating perceptual confidence from discrimination accuracy reveals no influence of metacognitive awareness on working memory. *Front. Psychol.* 7, 851
42. Rounis, E. *et al.* (2010) Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. *Cogn. Neurosci.* 1, 165–175
43. Fleming, S.M. and Daw, N.D. (2017) Self-evaluation of decision-making: a general Bayesian framework for metacognitive computation. *Psychol. Rev.* 124, 91–114
44. Barrett, A.B. *et al.* (2013) Measures of metacognition on signal-detection theoretic models. *Psychol. Methods* 18, 535–552
45. Rausch, M. *et al.* (2018) Confidence in masked orientation judgments is informed by both evidence and visibility. *Attention, Perception, Psychophys.* 80, 134–154
46. Rausch, M. *et al.* (2020) Cognitive modelling reveals distinct electrophysiological markers of decision confidence and error monitoring. *Neuroimage* 218, 116963
47. Rahnev, D. and Denison, R.N. (2018) Suboptimality in perceptual decision making. *Behav. Brain Sci.* 41, 1–107
48. Mueller, S.T. *et al.* (2008) Decision noise: an explanation for observed violations of signal detection theory. *Psychon. Bull. Rev.* 15, 465–494
49. Konishi, M. *et al.* (2020) Resilience of perceptual metacognition in a dual-task paradigm. *Psychon. Bull. Rev.* 27, 1259–1268
50. Maniscalco, B. *et al.* (2017) Limited cognitive resources explain a trade-off between perceptual and metacognitive vigilance. *J. Neurosci.* 37, 1213–1224
51. Rahnev, D. *et al.* (2015) Confidence leak in perceptual decision making. *Psychol. Sci.* 26, 1664–1680
52. Kantner, J. *et al.* (2019) Confidence carryover during interleaved memory and perception judgments. *Mem. Cogn.* 47, 195–211
53. Aguilar-Lleyda, D. *et al.* (2019) Confidence can be automatically integrated across two visual decisions. *PsyArXiv* Published online October 21, 2019. https://doi.org/10.31234/osf.io/3465b
54. Hauser, T.U. *et al.* (2017) Noradrenaline blockade specifically enhances metacognitive performance. *Elife* 6, e24901
55. Allen, M. *et al.* (2016) Unexpected arousal modulates the influence of sensory noise on confidence. *Elife* 5, e18103
56. Rhodes, M.G. and Castel, A.D. (2008) Memory predictions are influenced by perceptual information: evidence for metacognitive illusions. *J. Exp. Psychol. Gen.* 137, 615–625
57. Ferrigno, S. *et al.* (2017) A metacognitive illusion in monkeys. *Proc. R. Soc. B Biol. Sci.* 284, 20171541

58. Spence, M.L. *et al.* (2016) Computations underlying confidence in visual perception. *J. Exp. Psychol. Hum. Percept. Perform.* 42, 671–682
59. Fetsch, C.R. *et al.* (2014) Effects of cortical microstimulation on confidence in a perceptual decision. *Neuron* 83, 797–804
60. Boldt, A. *et al.* (2017) The impact of evidence reliability on sensitivity and bias in decision confidence. *J. Exp. Psychol. Hum. Percept. Perform.* 43, 1520–1531
61. de Gardelle, V. and Mamassian, P. (2015) Weighting mean and variability during confidence judgments. *PLoS One* 10, e0120870
62. Spence, M.L. *et al.* (2018) Uncertainty information that is irrelevant for report impacts confidence judgments. *J. Exp. Psychol. Hum. Percept. Perform.* 44, 1981–1994
63. Navajas, J. *et al.* (2017) The idiosyncratic nature of confidence. *Nat. Hum. Behav.* 1, 810–818
64. Gajdos, T. *et al.* (2019) Revealing subthreshold motor contributions to perceptual confidence. *Neurosci. Conscious.* 2019, niz001
65. Palser, E.R. *et al.* (2018) Altering movement parameters disrupts metacognitive accuracy. *Conscious. Cogn.* 57, 33–40
66. Kiani, R. *et al.* (2014) Choice certainty is informed by both evidence and decision time. *Neuron* 84, 1329–1342
67. Peters, M.A.K. *et al.* (2017) Perceptual confidence neglects decision-incongruent evidence in the brain. *Nat. Hum. Behav.* 1, 0139
68. Zylberberg, A. *et al.* (2012) The construction of confidence in a perceptual decision. *Front. Integr. Neurosci.* 6, 79
69. Koizumi, A. *et al.* (2015) Does perceptual confidence facilitate cognitive control? *Attention, Perception, Psychophys.* 77, 1295–1306
70. Zylberberg, A. *et al.* (2014) Variance misperception explains illusions of confidence in simple perceptual decisions. *Conscious. Cogn.* 27, 246–253
71. Forstmann, B.U. *et al.* (2016) Sequential sampling models in cognitive neuroscience: advantages, applications, and extensions. *Annu. Rev. Psychol.* 67, 641–666
72. Pleskac, T.J. and Busemeyer, J.R. (2010) Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychol. Rev.* 117, 864–901
73. Kiani, R. and Shadlen, M.N. (2009) Representation of confidence associated with a decision by neurons in the parietal cortex. *Science* 324, 759–764
74. Li, H.H. and Ma, W.J. (2020) Confidence reports in decision-making with multiple alternatives violate the Bayesian confidence hypothesis. *Nat. Commun.* 11, 1–11
75. Adler, W.T. and Ma, W.J. (2018) Comparing Bayesian and non-Bayesian accounts of human confidence reports. *PLoS Comput. Biol.* 14, e1006572
76. Adler, W.T. and Ma, W.J. (2018) Limitations of proposed signatures of Bayesian confidence. *Neural Comput.* 30, 3327–3354
77. Denison, R.N. *et al.* (2018) Humans incorporate attention-dependent uncertainty into perceptual decisions and confidence. *Proc. Natl. Acad. Sci. U. S. A.* 115, 11090–11095
78. De Martino, B. *et al.* (2013) Confidence in value-based choice. *Nat. Neurosci.* 16, 105–110
79. Drugowitsch, J. *et al.* (2016) Computational precision of mental inference as critical source of human choice suboptimality. *Neuron* 92, 1398–1411
80. Rahnev, D. *et al.* (2020) The confidence database. *Nat. Hum. Behav.* 4, 317–325
81. Kelemen, W.L. *et al.* (2000) Individual differences in metacognition: evidence against a general metacognitive ability. *Mem. Cogn.* 28, 92–107
82. Lee, A.L.F. *et al.* (2018) Cross-domain association in metacognitive efficiency depends on first-order task types. *Front. Psychol.* 9, 2464
83. Fitzgerald, L.M. *et al.* (2017) Domain-specific and domain-general processes underlying metacognitive judgments. *Conscious. Cogn.* 49, 264–277
84. Mazancieux, A. *et al.* (2020) Is there a G factor for metacognition? Correlations in retrospective metacognitive sensitivity across tasks. *J. Exp. Psychol. Gen.* 149, 1788–1799
85. Beck, B. *et al.* (2019) Metacognition across sensory modalities: vision, warmth, and nociceptive pain. *Cognition* 186, 32–41
86. Chiappe, D. and MacDonald, K. (2005) The evolution of domain-general mechanisms in intelligence and learning. *J. Gen. Psychol.* 132, 5–40
87. Baer, J. (2015) *Domain Specificity of Creativity*, Elsevier
88. Rouault, M. *et al.* (2018) Psychiatric symptom dimensions are associated with dissociable shifts in metacognition but not task performance. *Biol. Psychiatry* 84, 443–451
89. Wells, A. *et al.* (2012) Metacognitive therapy in treatment-resistant depression: a platform trial. *Behav. Res. Ther.* 50, 367–373
90. Moritz, S. *et al.* (2014) Sowing the seeds of doubt: a narrative review on metacognitive training in schizophrenia. *Clin. Psychol. Rev.* 34, 358–366
91. Klein, T.A. *et al.* (2013) Error awareness and the insula: links to neurological and psychiatric diseases. *Front. Hum. Neurosci.* 7, 14
92. Stephan, K.E. *et al.* (2009) Dysconnection in schizophrenia: from abnormal synaptic plasticity to failures of self-monitoring. *Schizophr. Bull.* 35, 509–527
93. Wixted, J.T. and Wells, G.L. (2017) The relationship between eyewitness confidence and identification accuracy: a new synthesis. *Psychol. Sci. Public Interest* 18, 10–65
94. Fleming, S.M. *et al.* (2012) Metacognition: computation, biology and function. *Philos. Trans. R. Soc. B Biol. Sci.* 367, 1280–1286
95. Koriat, A. (2006) Metacognition and consciousness. *Cambridge Handb. Conscious.* 3, 289–326
96. Shimamura, A.P. (2000) Toward a cognitive neuroscience of metacognition. *Conscious. Cogn.* 9, 313–323
97. Yeung, N. and Summerfield, C. (2012) Metacognition in human decision-making: confidence and error monitoring. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 367, 1310–1321
98. Rahnev, D. (2020) Confidence in the real world. *Trends Cogn. Sci.* 24, 590–591