OPEN MIND

Discoveries in Cognitive Science

an open access journal

Citation: Dayan, P. (2023). Metacognitive Information Theory. *Open Mind: Discoveries in Cognitive Science*, *7*, 392–411. https://doi.org/10.1162/opmi_a_00091

DOI:
https://doi.org/10.1162/opmi_a_00091

Supplemental Materials:
https://doi.org/10.1162/opmi_a_00091

Received: 24 October 2022
Accepted: 25 June 2023

Competing Interests: The authors declare no conflict of interests.

Corresponding Author:
Peter Dayan
dayan@tue.mpg.de

The MIT Press

REPORT

Metacognitive Information Theory

Peter Dayan[1,2]

[1]Max Planck Institute for Biological Cybernetics, Tübingen, Germany
[2]University of Tübingen, Tübingen, Germany

Keywords: metacognition, information theory, signal detection theory, scoring rules, m-ratio

ABSTRACT

The capacity that subjects have to rate confidence in their choices is a form of metacognition, and can be assessed according to bias, sensitivity and efficiency. Rich networks of domain-specific and domain-general regions of the brain are involved in the rating, and are associated with its quality and its use for regulating the processes of thinking and acting. Sensitivity and efficiency are often measured by quantities called meta–d' and the M-ratio that are based on reverse engineering the potential accuracy of the original, primary, choice that is implied by the quality of the confidence judgements. Here, we advocate a straightforward measure of sensitivity, called meta–$\mathcal{I}$, which assesses the mutual information between the accuracy of the subject's choices and the confidence reports, and two normalized versions of this measure that quantify efficiency in different regimes. Unlike most other measures, meta–$\mathcal{I}$-based quantities increase with the number of correctly assessed bins with which confidence is reported. We illustrate meta–$\mathcal{I}$ on data from a perceptual decision-making task, and via a simple form of simulated second-order metacognitive observer.

INTRODUCTION

The confidence that we apportion to our recollections, cognitions, decisions and actions can play a critical role in the preparations we make for success or failure; in determining whether we need to collect more external information or more samples of internal information before committing ourselves; in regulating the learning that we should do when outcomes are, or are not, as expected; and in communicating with others, for instance when engaging in collective decision-making (Bahrami et al., 2010; De Martino et al., 2013; Fleming, 2021; Fleming et al., 2012; Kepecs & Mainen, 2012; Nelson & Narens, 1990; Schulz et al., 2020). Confidence, as one of the simplest forms of higher-order or self-reflective assessment about one's own cognitive processes, has also been (sometimes controversially) influential in modern theories of awareness (Fernandez-Duque et al., 2000; Lau, 2022; Lau & Rosenthal, 2011). Furthermore, various impairments of metacognition are central in a number of psychiatric conditions, for instance, with possibly exorbitant requirements for confidence helping underpin excessive checking in forms of obsessive compulsive disorder; or substantial over-confidence helping reinforce the persistent apparently erroneous conclusions drawn by those suffering from delusional disorders (Hoven et al., 2019; Rouault et al., 2018; Seow et al., 2021; Sun et al., 2017). Various regions of the prefrontal cortex, anterior cingulate cortex, insular cortex and the precuneus have been implicated in making such judgements and using them to control our cognition (for a meta-analysis of a wealth of studies, see Vaccaro & Fleming, 2018).

It has therefore long been recognized that is critical to measure the nature and quality of confidence judgements. At stake are three, related quantities: *bias*, *sensitivity*, and *efficiency*

(Fleming & Lau, 2014; Galvin et al., 2003; Maniscalco & Lau, 2012; Nelson, 1984). For concreteness, consider a simple perceptual decision-making problem: judging whether a Gabor patch is tilted left (L) or right (R) of vertical. The sensory input α can abstractly be regarded as a noisy version of $d = -1$ (L) or $d = 1$ (R). On each trial, subjects report their decision (the 'action' a) about the tilt—this is often called a type 1 judgement—and also their degree of confidence in the rectitude or accuracy of that decision (the 'rating')—a type 2 judgement. For convenience, we consider the original report as coming from an 'actor', and the confidence judgement from a 'rater'; although these are, of course, the same individual (Schulz et al., 2023). The type 1 judgement is the topic of conventional signal detection theory (Green & Swets, 1966), with accuracy being quantified by such measures as type 1 sensitivity or d'. If the rating is interpreted as just the probability that the type 1 decision is correct (Pouget et al., 2016), then metacognitive bias measures the overall calibration of the rating—whether subjects tend to think that they are more or less accurate than they actually are. Of course, a subject could be metacognitively unbiased if she reported the correct overall probability of being correct on every trial, independent of the actual observation (like a well-calibrated, but useless, weather forecaster reporting on every day of the year, the overall mean probability of rain; Dawid, 1982). Thus, metacognitive sensitivity measures the adaptability of the rating to the actual rectitude on a trial-by-trial basis—an ideal rater would have perfectly predictive error monitoring, rating correctly on a trial-by-trial basis whether the type 1 decision is going to be proved correct or incorrect. However, metacognitive sensitivity is not the whole story—the rater has a particularly easy job if the type 1 action is generally correct—it would be hard for the rater to be incorrect. Thus metacognitive efficiency attempts to correct the sensitivity for the quality of inference. Of course, metacognitive bias also has an impact: a thoroughly metacognitively biased rater who declares themselves fully confident on every trial, even when she in fact errs, would necessarily be fully insensitive and inefficient.

It might seem obvious, at least to the Bayesian decision theorists amongst us, that sensible observers would use all the information available to make their type 1 choice on a trial ($a = 1$ if $P(d = 1|\alpha) > 0.5$), and the same information to make their type 2 rating ($P(d = 1|\alpha)$) about their type 1 choice. This would be bias-free, and would leave sensitivity and efficiency at maximal values given the decision-maker's perceptual capacities (d'). This would render nugatory the metacognitive measures. However, empirical findings do not accord with this expectation (for instance, it is impossible for the rating to be of a less than 50% chance of being correct; whereas subjects can actually be aware of upcoming errors before they occur (Gehring et al., 1993); also evident in signals that likely emanate from the anterior cingulate cortex (Botvinick et al., 2004; Carter et al., 1998; Dehaene et al., 1994; Kerns et al., 2004). Thus, there are various accounts in which, for instance, in a so-called second order model, the internal rater has access to both additional information after the type 1 decision has been made (for instance from so-called post-decisional information, which we later call γ, that has not been processed at the time that the decision is registered), and/or only a noisy internal report (β) of the information α that the actor used in making the type 1 decision in the first place (Fleming & Daw, 2017; Jang et al., 2012). The rater could also suffer from noise in their metacognitive judgement or report (Guggenmos, 2022, Shekhar & Rahnev, 2021). In cases such as this, it is possible to have metacognitive hypo- or hyper-sensitivity, and for the rater to predict errors.

When the type 2 decision is not a trivial function of type 1 information, much effort has gone into determining useful measures of metacognitive sensitivity and efficiency (Evans & Azzopardi, 2007; Ferrell & McGoey, 1980; Fleming & Lau, 2014; Galvin et al., 2003; Guggenmos, 2021; Kunimoto et al., 2001; Maniscalco & Lau, 2012; Nelson, 1984). One influential and attractive idea has been to imagine the type 1 decisions that the rater would have

been able to make, and assess the notional type 1 sensitivity of this rater (Maniscalco & Lau, 2012, 2014). This value is called meta–d', and has the attractive characteristic of being directly comparable to the actual type 1 sensitivity. Meta–d' can be assessed by fitting the actual confidence statistics to the confidence statistics that would have been predicted by the imaginary type 1 choices of the rater. It is then possible to create a metacognitive efficiency measure that adjusts for the underlying ease of the decision-making problem by comparing meta–d' to d' either subtractively (meta–d' — d') or divisively (the so-called M-ratio, which is meta–d'/d').

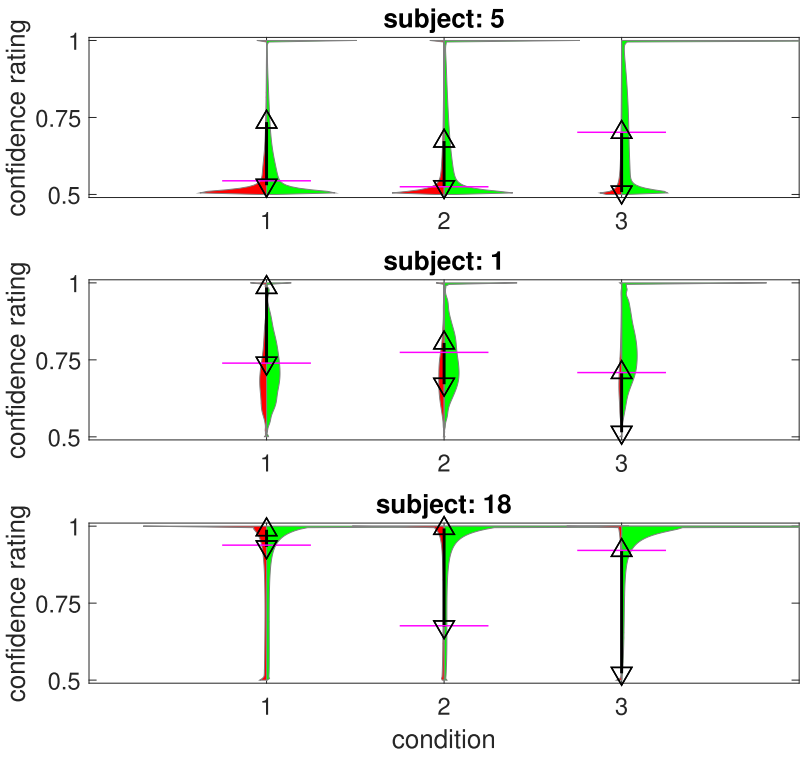
Meta–d' and the M-ratio are widely used as measures of meta-cognitive effectiveness (Barrett et al., 2013). However, along with some obvious assumptions (such as that ratings are monotonic in expected accuracy), they have some less desirable characteristics, including remaining dependency of the M-ratio on type 1 performance (Guggenmos, 2021, at least some aspects of which are, as noted above, inevitable), and on metacognitive bias (Xue et al., 2021, which is arguably less so). Here, along with the common observation that there is no reason to expect subjects' empirical confidence judgements to fit an assumed type 1 decision-process exactly, which means that the assessment of metacognitive efficiency could be inaccurate, we focus on the fact that the M-ratio does not take explicit account of the number of levels of confidence rating that subjects might be able to provide. A rater who can make a fine discrimination between being correct within the intervals [80, 85)% or [85, 90)% might reasonably demand to be considered more sensitive (and more efficient) than one with a single rating 'bucket' for the whole range [80, 90)%. Given a rater whose confidence is perfectly consistent with a type 1 decision, this excess discriminability will normally have no benefit from the perspective of meta–d'.

Here, we introduce and explore a natural alternative to meta–d' and the M-ratio, namely meta–$\mathcal{I}$ and two forms of a meta–$\mathcal{I}$-ratio (called meta–$\mathcal{I}^r_1$ and meta–$\mathcal{I}^r_2$), which are based on the mutual information between the rectitude of the actor and the confidence ratings. The mutual information is straightforward to compute for conventional rating buckets, makes fewer assumptions about the ratings, other than that they are distinct and, ideally, suitably predictive of differences in accuracy, and increases naturally with the granularity of the ratings. The mutual information is related to measures that are based on the correlation between accuracy and confidence (Nelson, 1984), although it is, for instance, completely agnostic to any bias. We first illustrate meta–$\mathcal{I}$-based measures on confidence data from Shekhar and Rahnev (2021). Then, to examine their properties in detail, we use a simple realization of a second-order rater (Fleming & Daw, 2017; Guggenmos, 2022; Jang et al., 2012; Mamassian & de Gardelle, 2022; Schulz et al., 2023), for which we can precisely unpick the nature of metacognitive sensitivity.

META–$\mathcal{I}$

Consider a simple perceptual decision-making task such as that reported in Shekhar and Rahnev (2021). Here, on each of 2800 trials t, participants saw for just 100 ms a noisy Gabor patch (of one of three different contrasts, defining three *conditions*) that was tilted either to the left or right of vertical (we write this as $d^t = \pm 1$), and used a single scale to report the direction of the tilt ($a^t = \pm 1$) and a continuous confidence rating (c^t) about the accuracy of their choices (whose true value is $r^t = d^t \times a^t$). Confidence reporting in this experiment restricted c^t to being between 0.5 and 1.

To illustrate the issues for measuring the quality of metacognition, Figure 1 shows violin plots of the distributions of confidence reports for three selected subjects for incorrect (red) and correct (green) choices and for the three contrast conditions (1 is hardest; 3 is easiest). Subject 5 is biased to report low confidence; subject 18 to report high confidence; subject 1 is in the middle. We can see the reduction in incorrect responses with higher contrast



sub	con	d'	md'	mR	m$\mathcal{I}$[2]	m$\mathcal{I}$[3]	m$\mathcal{I}$[10]	m$\mathcal{I}_1^r$[2]	m$\mathcal{I}_1^r$[3]	m$\mathcal{I}_1^r$[10]
5	1	0.7	0.9	1.3	0.044	0.050	0.064	1.372	1.541	1.980
	2	1.1	1.1	1.1	0.057	0.068	0.081	0.984	1.181	1.412
	3	2.1	2.0	0.9	0.094	0.118	0.136	0.775	0.978	1.128
1	1	1.1	0.5	0.5	0.019	0.023	0.036	0.329	0.415	0.639
	2	1.6	0.9	0.6	0.031	0.036	0.049	0.323	0.375	0.513
	3	2.4	1.8	0.7	0.067	0.080	0.102	0.538	0.643	0.818
18	1	0.9	0.6	0.7	0.019	0.022	0.037	0.435	0.502	0.858
	2	1.7	1.0	0.6	0.029	0.039	0.052	0.283	0.381	0.512
	3	2.8	1.6	0.6	0.030	0.045	0.058	0.248	0.370	0.477

Figure 1. Confidence and accuracy for three selected subjects ('sub'). Each plot shows the distribution of confidence reports for incorrect (red) and correct (green) choices for each of the three contrast conditions ('con'). The total area of the violin plots is normalized, so the overall accuracy is evident in the sizes of the green areas. The magenta bars show the division of confidence ratings into two bins that approximately maximize meta–$\mathcal{I}$; the lower and upper triangles show the same for three bins. The table provides the numerical values of the various sensitivity and efficiency measures for these subjects. md' is meta–d', mR is the M-ratio, m$\mathcal{I}[n]$ is meta–$\mathcal{I}$ for n, approximately optimally-positioned, confidence bins, m$\mathcal{I}_1^r[n]$ is meta–$\mathcal{I}_1^r$ for those same bins. Data from Shekhar and Rahnev (2021).

(i.e., higher condition number); but also additional facets such as the spikes of very high confidence reports. Since confidence should provide information about accuracy, measures of meta-cognitive sensitivity report how closely related are c^t and r^t. In terms of the plots in Figure 1, we would seek the mass of the green distributions to be higher than those of the red ones, with higher confidence ratings when the answer is actually correct.

As mentioned above, there is a wide variety of such measures, some of which are based on process models of the way that participants make decisions and rate confidence (e.g., Desender et al., 2021; Maniscalco & Lau, 2012; Shekhar & Rahnev, 2021), whereas others are agnostic to the process by which confidence judgements are made, and depend on some form of correlation between accuracy and confidence (Nelson, 1984). In both cases, raw assessments are influenced by the absolute accuracy, since, for instance, if the decision-making task is very

easy, there is little uncertainty to which confidence could be sensitive. The table below the figure indicates d', meta–d' (written as md') and the M-ratio (mR). Here, these quantities were calculated using maximum likelihood fitting routines from Maniscalco and Lau (2012, 2014), in which the parameters of a naïve first order Bayesian rater are fit so that the distribution of its confidence ratings match those of each subject. From the M-ratio, we can see that, in this case, the order of the efficiency of these subjects is opposite to the order of their metacognitive bias.

Figure 2A and B show meta–d' and the M-ratio for all the subjects in Shekhar and Rahnev (2021), in the three contrast condition. The subjects are sorted differently in each figure in decreasing order of the sensitivity (Figure 2A) or efficiency (Figure 2B) for the most difficult condition (blue). We see that both measures tend to decrease together for all the contrast conditions, confirming past observations that there something generalizable about sensitivity and efficiency, at least for such closely related problems. Figure 2A also shows the dependence of meta–d' on d': as noted, these values are distinctly greater for the higher contrast conditions. Figure 2B shows that this characteristic is largely abolished for the M-ratio, in which the rater's meta–d' is normalized by the actor's d'.

The measure meta–d' (along with others mentioned above) is a model-based measure of sensitivity, in that one imagines that the confidence reports are the first order judgements for an actor with some particular, parametrized characteristics. By contrast, meta–$\mathcal{I}$ is a model-agnostic measure of metacognitive sensitivity which quantifies the mutual information between r^t and c^t. Take the case that confidence is discrete (in Shekhar and Rahnev (2021), it is measured in 1/1000$^{\text{ths}}$) and that we had been able to measure the full joint distribution of

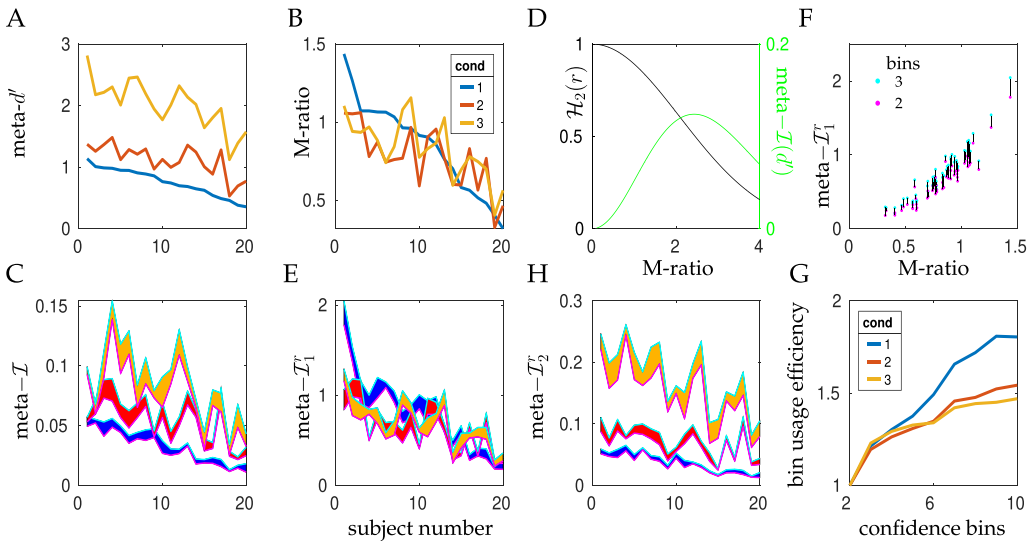


Figure 2. Metacognition in Shekhar and Rahnev (2021). (A) meta–d' across the 20 subjects for the three contrast conditions (in order of contrast, blue, red, yellow), with the subjects sorted by the lowest contrast. (B) The M-ratio for the subjects, again sorted by the value in the lowest contrast condition. (C) meta–$\mathcal{I}$ for the subjects in the same sort order as in (A) for two (lower envelope in magenta) or three (upper envelope in green) confidence bins with optimized thresholds. The lozenges fill the areas between lower and upper envelope in the colours for the condition. (D) The two normalizers for assessing efficiency: meta–$\mathcal{I}(d')$ (green; right axis, used to define meta–$\mathcal{I}_1^r$) and $\mathcal{H}_2(r)$ (black; left axis, used to define meta–$\mathcal{I}_2^r$) as a function of the actor's d'. (E) meta–$\mathcal{I}_1^r$ for the subjects in the same sort order as in (B) with the same plotting conventions as in (C). (F) The relationship between the M-ratio and meta–$\mathcal{I}_1^r$ for the 60 combinations of subjects and conditions, for two (magenta) and three (green) confidence bins, joined by thin black lines for clarity. (G) The average across subjects of the ratio between meta–$\mathcal{I}$ (or equivalently meta–$\mathcal{I}_1^r$ or meta–$\mathcal{I}_2^r$) for 2 ... 10 confidence bins to meta–$\mathcal{I}$ for just 2 confidence bins, for the three conditions. This shows that the subjects are, on average, able to use multiple bins to some good effect. (H) meta–$\mathcal{I}_2^r$ for the subjects in the same sort order as in (B) with the same plotting conventions as in (C).

rating and confidence, $P(r, c)$, with $P(r) = \Sigma_c P(r, c)$; $P(c) = \Sigma_r P(r, c)$. Then, the mutual information is the difference between two entropies (measured, for convenience, in bits). One entropy, which we write as $\mathcal{H}_2(r)$, is the overall uncertainty about the accuracy of the actor. For binary choice, this quantity varies between 0 bits, if the actor is perfectly accurate, as $d' \to \infty$ (or indeed if the actor is perfectly inaccurate; always getting the answer wrong) and 1 bit, if the action a is completely uncorrelated with the truth d, which happens as $d' \to 0$. The second entropy, $\mathcal{H}_2(r|c)$ is the weighted average uncertainty about the accuracy that remains after observing the confidence rating c, where the weights come from the probability of seeing that rating c. The confidence judgement is very sensitive and efficient if most of the initial uncertainty about the accuracy is removed by the rating, making this last term near 0.

More formally,

$$\text{meta–}\mathcal{I} = \mathcal{H}_2(r) - \mathcal{H}_2(r|c) \quad \text{where} \tag{1}$$

$$\mathcal{H}_2(r) = h_2[P(r)] \quad \text{is the entropy of the accuracy,} \tag{2}$$

$$\mathcal{H}_2(r|c) = \sum\nolimits_c P(c) h_2[P(r|c)] \quad \text{is the conditional entropy of accuracy given confidence;} \tag{3}$$

$$h_2[P(x)] = \sum\nolimits_{x=0;1} -P(x) \log_2 P(x) \quad \text{is the entropy of a Bernoulli random variable} \tag{4}$$

Mutual information is symmetric, so one can also write meta–$\mathcal{I} = \mathcal{H}_2(c) - \mathcal{H}_2(c|r)$. As is sometimes common, one could condition all these quantities on the action, and so report response-specific meta–$\mathcal{I}(a = 1)$ and meta–$\mathcal{I}(a = -1)$.

Table 1 provides an illustration of the way that one can calculate mutual information. Here, we show the four combinations of correct ($r = 1$) and incorrect ($r = 0$) and *high* ($c = h$) and *low* ($c = l$) confidence for condition 3 for subject 5 (see the rightmost violin plot in the first row of Figure 1). We binarized the subject's nearly continuous report of confidence at the optimal point shown by the magenta bar in the figure (i.e., a threshold of 0.72). In this case, the probability of being correct is $P(r = 1) = 855/998$, with an entropy of $\mathcal{H}_2(r) = 0.593$ bits; the probability of high and low confidence are $P(c = h) = 420/998$ and $P(c = l) = 578/998$ respectively; the conditional probability of being correct given *high* confidence is $P(r = 1|c = h) = 415/420$ with an entropy of $h_2[P(r|c = h)] = 0.093$ bits; and the conditional probability of being correct

Table 1. Calculation tableau for meta–$\mathcal{I}$ for condition 3 for subject 5 (see also Figure 1). Here, we see the prevalence of the four combinations of being correct ($r = 1$) or incorrect ($r = 0$) and having high ($c = h$) or low ($c = l$) confidence, dividing the subject's judgement at the single threshold of 0.72 shown by the magenta line in Figure 1

rectitude	confidence	prevalence
$r = 0$	$c = l$	138
$r = 0$	$c = h$	5
$r = 1$	$c = l$	440
$r = 1$	$c = h$	415

given *low* confidence is $P(r = 1|c = l) = 440/548$ with an entropy of $h_2[P(r|c = l)] = 0.793$ bits. Thus the full mutual information is

$$\text{meta–}\mathcal{I} = \mathcal{H}_2(r) - P(c = h) \times h_2[P(r|c = h)] - P(c = l) \times h_2[P(r|c = l)]$$

$$= 0.593 \qquad -\frac{420}{998} \times 0.093 \qquad -\frac{578}{998} \times 0.793$$

$$= 0.094$$

as in the figure.

Figure 2C generalizes this to show meta–$\mathcal{I}$ for two ways of turning the (nearly) continuous measure of confidence that the subjects reported into a set of mutually exclusive bins.[1] The lower (magenta) border of the lozenge for each condition (distinguished by the fill colour) is the result of choosing the best binarization of confidence (something that Shekhar and Rahnev (2021) explored explicitly). The short horizontal magenta lines on the violin plots of Figure 1 show where this binary separation falls for those three subjects—trying to separate green and red masses vertically. These thresholds are ultimately an expression of the bias in confidence reporting of the subjects—we see their levels roughly reflecting how the subjects employ the confidence scale. The upper (green) border of each lozenge shows the case of three optimized levels of confidence arising from the two thresholds shown as lower and upper triangles in Figure 1. First, note that meta–$\mathcal{I}$ values are (by construction) higher for the extra bins—a phenomenon we explore further below. Second, the subjects are sorted as in Figure 2A (by meta–d' in the lowest contrast condition); however the lozenge for this condition is almost monotonic; and the lozenges for the higher contrasts have a similar degree of monoticity to those in meta–d', suggesting that meta–$\mathcal{I}$ is somewhat consistent with meta–d'. Along with this, we see that meta–$\mathcal{I}$ also increases with d'—although we will later qualify this finding—it arises here partly because of the rather modest levels of the actors' d's in this study (with most M-ratios being less than 1).

meta–$\mathcal{I}$ is a measure of metacognitive *sensitivity*. As for the relationship between meta–d' and the M-ratio, measuring metacognitive *efficiency* requires normalizing for a quantification of potentially available information about confidence. Guggenmos (personal communication) thus suggested taking the analagous step of calculating a meta–$\mathcal{I}$-ratio by normalizing meta–$\mathcal{I}$. One possible normalizer would be a quantity we write as meta–$\mathcal{I}(d')$ that would arise as the value of meta–$\mathcal{I}$ for a first-order rater following a conventional signal detection theory analysis based on the actor's d' where

$$\alpha \sim \mathcal{N}\left(d, \frac{4}{(d')^2}\right) \qquad a = \text{sign}(\alpha) \qquad c = \frac{1}{1 + \exp(-d'|\alpha|)}$$

The green curve in Figure 2D shows meta–$\mathcal{I}(d')$ as a function of d'. For relatively low values of d', as seen in Shekhar and Rahnev (2021), this increases with d'. However, for large d', it decreases again, since meta–$\mathcal{I}$ is bounded above by the entropy of the accuracy $\mathcal{H}_2(r)$—and as d' rises, the actor becomes increasingly accurate, and so this entropy decreases.

[1] See Shekhar and Rahnev (2021) for their analysis of the quantization of the continuous confidence report. They also performed a sophisticated examination of different models of metacognitive rating based on this (notably suggesting the influence of a particular sort of noise). However, for our present purpose of analyzing our model-free construct meta–$\mathcal{I}$, we will make a simple comparison with a binarized estimate of meta–d'.

This manoeuvre of normalizing meta–$\mathcal{I}$ by meta–$\mathcal{I}(d')$ parallels the M-ratio's use of d' itself to normalize meta–d'. It captures the inability of a first order actor with poor perceptual abilities to judge confidence well; and the consequences for metacognitive sensitivity of the lack of variability in the rectitude of an actor with excellent perceptual abilities. We write meta–$\mathcal{I}_1^r$ = meta–$\mathcal{I}$/meta–$\mathcal{I}(d')$, and show it in Figure 2E for the two and three confidence bins of Figure 2C, but now sorted by the M-ratio of the lowest contrast condition (i.e., the sort used in Figure 2B). As for the M-ratio, we see that this normalization has more nearly equated the estimates of meta-cognitive efficiency for the three conditions, to a roughly equivalent degree to the M-ratio.

Figure 2F shows the relationship between the M-ratio and meta–$\mathcal{I}_1^r$ for all subjects and all conditions for the case of two (magenta) and three (green) confidence bins (with the cases joined by vertical lines). We can see that there is a very close relationship between the M-ratio and meta–$\mathcal{I}_1^r$, at least in this regime of actors and raters, confirming the impression from comparing Figures 1B and 1D.

How, though, should we think about the fact that there are apparently different values of meta–$\mathcal{I}$ for different numbers of confidence bins? All else being equal, a rater that can accurately distinguish a larger number of levels of accuracy should reasonably be considered to be more metacognitively sensitive and efficient—since this rater can offer a finer perspective on the chance of failure. Equivalently, meta–$\mathcal{I}$ will benefit from the deblurring of the ratings that occurs when they are split into more levels, at least provided that these levels are used well. This is not a property of meta–d' or the M-ratio—the main consequence of increasing the granularity of the confidence report is to affect the fitting process for estimating the rater's equivalent d'—it has no direct bearing on that version of sensitivity or efficiency.

To assess the consequence of increasing granularity, we evaluated the average across the subjects in Shekhar and Rahnev (2021) of the ratio of meta–$\mathcal{I}$ for between two and ten confidence bins and for just two bins. Here, we approximately optimized the thresholds on a subject- and condition-specific basis (Figure 2G). From the increase with the number of bins, it is apparent that the subjects are able on average to report confidence at at relatively fine granularity—particularly in the most difficult (blue) contrast condition—but that this ultimately saturates (with many fewer than the ~500 confidence bins of the experimental report). One wrinkle here is that we calculated the efficiency normalizer, meta–$\mathcal{I}(d')$, assuming continuous confidence judgements can be made by a first-order rater (i.e., with an infinite number of correctly-employed confidence bins). This is reasonable, because this estimate is based on a model that allows calculation to arbitrary accuracy. However, it could be questioned as a comparator, and it would also be possible to normalize by a version of meta–$\mathcal{I}(d', b)$ that uses to optimal effect the same number (b) of confidence bins as the empirical rater.

Figure 2H reports the result of normalizing meta–$\mathcal{I}$ by the theoretical upper bound $\mathcal{H}_2(r)$ we mentioned above rather than meta–$\mathcal{I}(d')$. We call the resulting measure of metacognitive efficiency meta–$\mathcal{I}_2^r$. $\mathcal{H}_2(r)$ is shown as a function of d' in the black curve in Figure 2D. This also accounts well for the fact that, for high d' for the actor, metacognitive sensitivity cannot be high, since, as we noted, there is little entropy in $\mathcal{H}_2(r)$ to reduce by $\mathcal{H}_2(r|c)$. However, in cases such as the second order model we consider later in which the rater can have access to much better information than the actor, it allows us to assess the efficiency in absolute term. The data in Figure 2H suggest that this regime is not relevant for the data in Shekhar and Rahnev (2021), in that the raters appear to be generally rather worse than the actors.

A final issue with information theoretic measures concerns estimation of entropies and conditional entropies. The mutual information associated with continuous variables, such as

confidence in some experiments, is known to be hard to estimate, because of biases, and so care is necessary (Kozachenko & Leonenko, 1987; Paninski, 2003; Panzeri & Treves, 1996; Witter & Houghton, 2021). Biases are typically weaker for discrete variables, which are employed in many experiments on confidence. Here, we use randomized or exact permutation methods as a simple way to correct for biases.

A SECOND-ORDER DECISION-MAKER

In order to examine meta–$\mathcal{I}$ and meta–$\mathcal{I}^r_{\{1,2\}}$ in more detail, we turned to a simulation which allows us to abstract the relevant factors away from the noise associated with the ratings of individuals. We simulate choices and ratings from a simple, realized form of a second-order decision-making (Fleming & Daw, 2017; Jang et al., 2012; Mamassian & de Gardelle, 2022). On any trial, the actor and rater collectively receive three Gaussian distributed signals that bear on a true underlying quantity $d = \pm 1$ (Figure 3A–C); the actor generates a binary choice a (Figure 3A); and the rater generates one of a number of discrete confidence values c (Figure 3E and 3H). Here:

$$\alpha = d + A\epsilon_\alpha \quad \text{is the primary decision variable with type-1 sensitivity } d' = 2/A \tag{5}$$

$$\beta = \alpha + B\epsilon_\beta \quad \text{allows the rater partial insight into the basis of the actor's decision} \tag{6}$$

$$\gamma = d + G\epsilon_\gamma \quad \text{provides the rater with independent or unique information about } d \tag{7}$$

where $\{\epsilon_\alpha, \epsilon_\beta, \epsilon_\gamma\}$ are independent standard $\mathcal{N}(0, 1)$ distributed random variables. We assume that $d = \pm 1$ with equal probability, and that the actor is unbiased, and so makes a decision based on the sign of α, with $a = -1$ if $\alpha < 0$, $a = 1$ if $\alpha > 0$ and is indifferent if $\alpha = 0$. The rater bases its choice on the observation of the action a and the random variables β and γ, whose combination arranges for potentially partial correlation between the actor's and rater's information about the true state of the stimulus, d, as in the standard second order model. Here, the rater's confidence $c = P(a = d|a, \beta, \gamma)$ is the probability that the actor's action a was correct given all the information that the rater possesses.

Although for didactic convenience, the realization of the information structure relating actor and rater is different here from the canonical stochastic detection and retrieval model of Jang et al. (2012), the ultimate statistical relationships are the same; we provide a translation in the Supplement. Thus, briefly, and as discussed at length in Fleming and Daw (2017), this model has various natural limits of metacognitive interest. In the case that $B \to 0$; $G \to \infty$, the rater has exactly the same information as the actor, and so could act like the naive Bayesian decision theorist above. Then, meta–d' would be the same as d', and the M-ratio would be 1. In the opposite case, $B \to \infty$; $G \to 0$, the rater would have perfect knowledge about the rectitude of the actor's choice (based on the equivalent of perfect post-decisional information, also known as a confidence 'boost'; Mamassian and de Gardelle (2022)) and so could be as sensitive as it is possible to be. If $B \to \infty$; $G \to \infty$, then the rater has no specific information about the basis of the actor's choice on a trial, and so, like the incompetent weather forecaster, could do no better than reporting the overall expected accuracy on each trial (which here is $\Phi(d'/2)$, where Φ is the cumulative distribution function for a normal distribution). In non-limiting cases, β provides the rater with information about the data on which the actor based their choice, which she has to combine with her private, e.g., post-decisional, information about d.

It is didactically convenient to consider response-specific confidence, although here, the symmetry of the problem means that all the measures will be the same for $a = -1$ and $a = 1$.

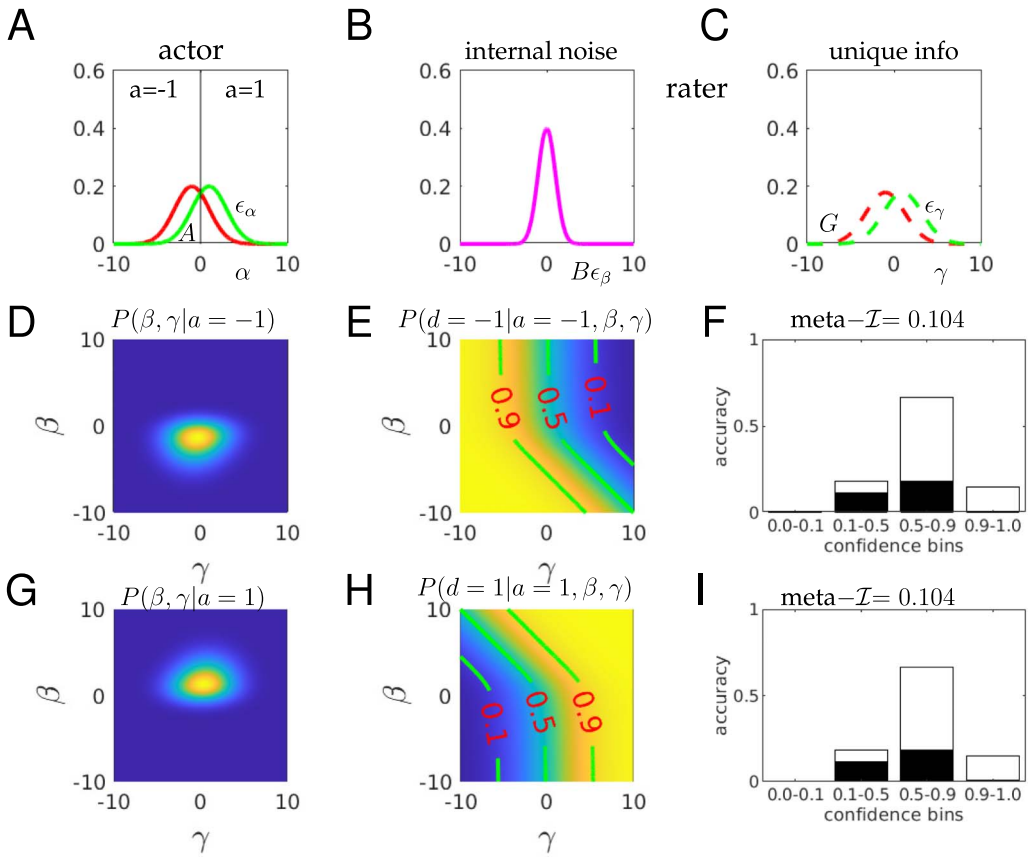


Figure 3. Mutual information calculations for a realized second-order model. (A) the actor observes a signal $\alpha = d + (2/d')\epsilon_\alpha$ (red for L: $d = -1$; green for R: $d = +1$) and makes an unbiased decision a for $d = \pm 1$ at the boundary $\alpha = 0$. (B, C) The rater receives two pieces of information: $\beta = \alpha + B\epsilon_\beta$; $\gamma = d + G\epsilon_\gamma$ where all ϵ are standard $\mathcal{N}(0, 1)$ and independent. γ is called unique since it contains information about d that is not shared with the actor. Here $d' = 1$; $B = 1$; $G = \sqrt{5}$. (D) The density $P(\beta, \gamma | a = -1)$ slightly favours the lower left quadrant, but with substantial noise. The distribution integrates to 1; color scale not shown for convenience. (E) The conditional probability $P(d = -1 \mid a = -1, \beta, \gamma)$ is the accuracy afforded by the rater's information set $(a = -1, \beta, \gamma)$. If $\beta, \gamma \ll 0$, then the decision $a = -1$ is likely to be true. The contour lines show the boundaries where this objective confidence crosses the values shown—the enclosed regions are where objective confidence ratings would be provided. (F) If we consider the regions of β, γ that define these bins of confidence, we can assess the expected accuracy—defined by the combination of the probability of ending up in one of the confidence bins $c(a = -1, \beta, \gamma)$ and the chance of being correct (white) or incorrect (black) in that bin. The mutual information $\mathcal{I}(r, c)$ between being correct and c (given $a = -1$) is 0.104 bits. (G–I) The same as (D–F) except for the case that $a = 1$. Since the problem is symmetric, this is essentially the same as for $a = -1$.

Figure 3D and 3E show the two critical quantities that govern confidence judgements for $a = -1$ (for the case that $d' = 1$; $B = 1$, $G = \sqrt{5}$). First, Figure 3D shows the posterior density that the rater will observe β and γ given that $a = -1$. These values slightly favour the lower left quadrant ($\beta < 0$; $\gamma < 0$), since $a = -1$ implies that the actor saw $\alpha < 0$. Note that this preference is stronger as a function of β than γ; this is because B is quite small, and we know that $\alpha < 0$ (since $a = -1$). Second, Figure 3E shows the optimal confidence $P(d = -1 | a = -1, \beta, \gamma)$ that the rater would have about $d = -1$ given all the information in her possession. The plot shows contours at {0.1, 0.5, 0.9} to indicate more precisely the shape of this distribution. Coarsely, if β is very negative, which, because B is small, likely means that α was very negative, then the rater is rather confident that $a = -1$ is correct, unless γ is very large and positive, to counteract this. The slopes of the contours for negative β largely reflect the relative information about d in α and γ. If β is very positive, then since $a = -1$, it can only be that α is very close to 0, and so the rater has to rely on γ, implying that the contours run largely perpendicular to the γ axis.

Figure 3F shows the consequence for the confidence ratings. Here, we consider the four bins implied by the contours in Figure 3E): {[0, 0.1), [0.1, 0.5), [0.5, 0.9)[0.9, 1.0]}. These bins were chosen to keep them separated on the plot; we consider issues of the nature of the bins later. The total height of each bar integrates the total probability mass (from Figure 3D) that ends up in each of the regions delineated in Figure 3E. This quantifies the fraction of confidence reports that will end up in each confidence bin. For each of these confidence reports, the actor could be correct or incorrect; we show the expected proportion of correct reports in white; and incorrect reports in black. If the confidence bins were very narrow, then since all calculations are probabilistically correct, the relative heights of black and white parts of a bar would be given by just the confidence level associated with this bar (since this is exactly what the confidence quantifies). However, since the confidence regions are rather wide, we have to calculate a weighted average, where the weights are purely determined by the probability mass in Figure 3D and the quantity that is averaged is the precise confidence in Figure 3E. Thus, for instance, the mean accuracy in the [0.5, 0.9) bin is slightly less than the centre of this internal (0.75). In this instance, we can calculate meta–d' = M-ratio = 1.7 based on the statistics in these confidence bins.

Figure 3G–H show exactly the same as Figure 3D–F, but for the case that $a = 1$ instead. The distributional plots are mirror symmetric, favouring positive rather than negative values of β, γ. The confidence values in Figure 3H are exactly the same as in Figure 3F, since this rater is just as good for $a = 1$ as for $a = -1$.

We now consider meta–$\mathcal{I}$ and meta–$\mathcal{I}^r_{\{1,2\}}$ for this case. First, the actor is 69% accurate (with $d' = 1$), making the unconditional entropy $\mathcal{H}_2(r|a = -1) = 0.89$ bits. Second, $P(c|a = -1)$ is the total height of the bars in Figure 3F for confidence rating $c \in \{[0, 0.1), [0.1, 0.5), [0.5, 0.9)[0.9, 1.0]\}$, and $P(r|c, a = -1)$ is the ratio between the black and white portions of those bars. The individual entropy terms for the bars in Figures 3F (defined by $\mathcal{H}_2[P(r|c \in \text{bin}_i, a = -1)]$) are 0.42, 0.95, 0.84, 0.34 bits respectively for the four bins, making the total remaining uncertainty about the accuracy as $\mathcal{H}_2(r|c, a = -1) = 0.79$. This leaves meta–$\mathcal{I} = 0.89 - 0.79 = 0.1$ bit, and, since meta–$\mathcal{I}(d' = 1) = 0.052$ bits, we have meta–$\mathcal{I}^r_1 = 2$. Here, although the rater is therefore more accurate than the actor (as similarly reflected by the M-ratio), her absolute efficiency meta–$\mathcal{I}^r_2 = 0.12$ is rather low, since the rater's unique information γ is subject to quite some noise, with G being large.

Figure 4 shows the same as Figure 3, but for the case that the rater enjoys a much greater amount of unique, post-decisional, information (with $G = \sqrt{0.5}$; Figure 4C). Figure 4D now shows bimodality, since γ is only likely to be near $d = \pm 1$, and less likely to take a value near 0. The mode associated with $\gamma = -1$ has a greater mass than that for $\gamma = 1$, since $a = -1$. The conditional distribution in Figure 4E now shows a much starker contrast—with the highly accurate γ being the main determinant of the confidence in the actor's choice (so if $\gamma > 0$, then the rater is rather confident that the actor erred). Figure 4F shows the consequence of this for the rating buckets. Now, the extreme values are much more likely—and are duly more pure in the sense that the rater can be rather sure about the rectitude of the actor. Here, meta–d' = M-ratio = 4.5, showing the benefit of the well-informed rater. Again, the case for $a = 1$ (Figure 4G–H) is the mirror image of the case for $a = -1$.

If we carry out the same calculations of meta–$\mathcal{I}$ and meta–$\mathcal{I}^r_{\{1,2\}}$ for this case, we observe that the individual entropy terms for the bars in Figure 4F are 0.15, 0.84, 0.82, 0.11 respectively. However, the fact that most of the weight in the average is on the outer two bars means that the total remaining uncertainty about the accuracy as $\mathcal{I}_2(r|c, a = -1) = 0.27$. This makes meta–$\mathcal{I} = 0.62$ bits, and meta–$\mathcal{I}^r_1 = 12$. In this regime, the M-ratio and meta–$\mathcal{I}^r_1$ diverge.

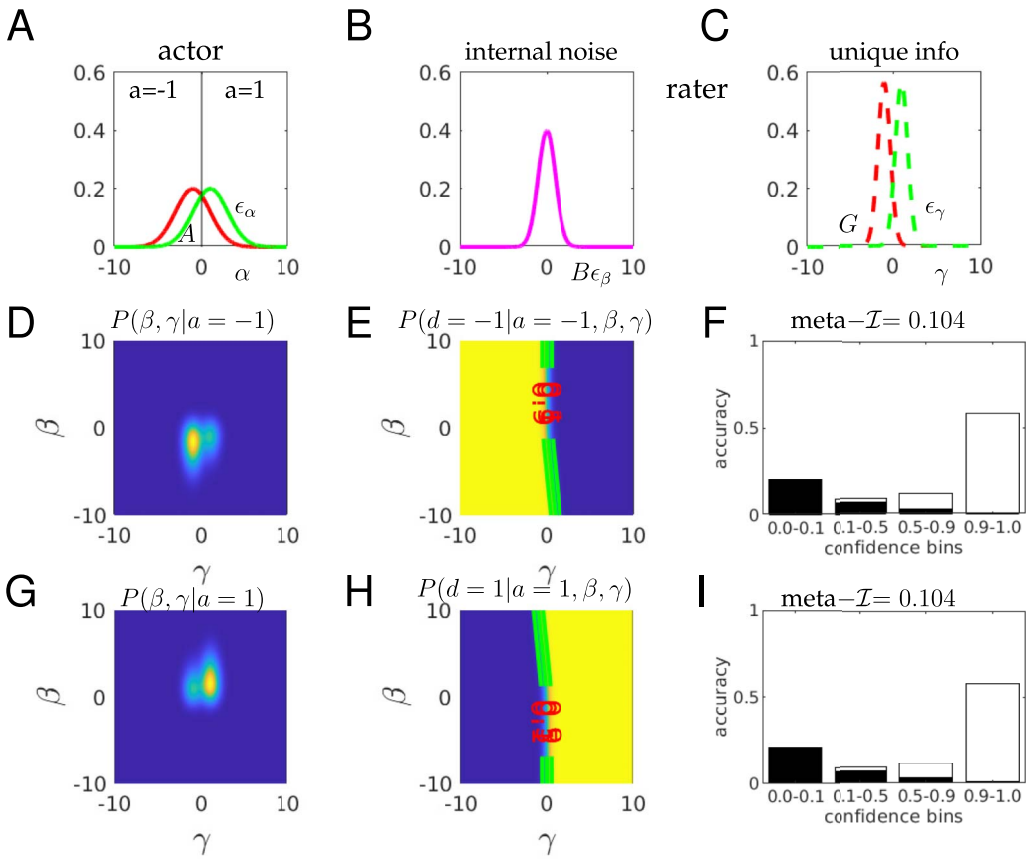


Figure 4. Mutual information calculations for a realized second-order model. This figure is the same as Figure 3, except that the standard deviation of the rater's unique information γ is $G = \sqrt{0.5}$. The bimodal distributions in (D, G) come from the two narrow possibilities for γ, with the weight on γ being near to -1 being higher in (D), because $a = -1$ there. Now the confidence contours are defined almost exclusively by γ (the near vertical arrangements in E and H); and the accuracy bins are nearly exclusively one colour (when the rater's confidence is 0–0.1, the actor is almost always incorrect.

However, the actor's performance is not a good yardstick for the rater, since the rater has substantially more information. Thus, meta–$\mathcal{I}_2^r = 0.7$ is a more useful measure of the high absolute efficiency of the rater, which reflects the high signal to noise ratio of the rater's unique information, with G being small.

Figure 5 compares the various metacognitive sensitivity and efficiency measures for various values of the actor's type 1 sensitivity d', and for different qualities of the unique information of the rater G^2. Here, as in Figures 3 and 4, $B = 1$. By contrast with the earlier figures, however, the rater optimally deployed four confidence bins.

These plots cover the two regimes discussed earlier. In one, where G^2 is not too small, the rater is at least co-dependent on the information that the actor used. Here the M-ratio (B) and meta–$\mathcal{I}_1^r$ (D) largely agree (although, as we will see later, meta–$\mathcal{I}_1^r$ correctly exploits extra confidence bins). However, in the other regime, where the rater is mostly dependent on its own source of information (G^2 is small), and the actor's performance is poor, then both the M-ratio and meta–$\mathcal{I}_1^r$ diverge. Here, the absolute efficiency, meta–$\mathcal{I}_2^r$ (E), of the rater is more relevant. Indeed, we can see that meta–$\mathcal{I}_2^r$ is largely constant as a function of d' for very low G^2. This property is shared by meta–d'; however meta–d' lacks an appropriate scale (since the actor's d' is not an appropriate baseline).

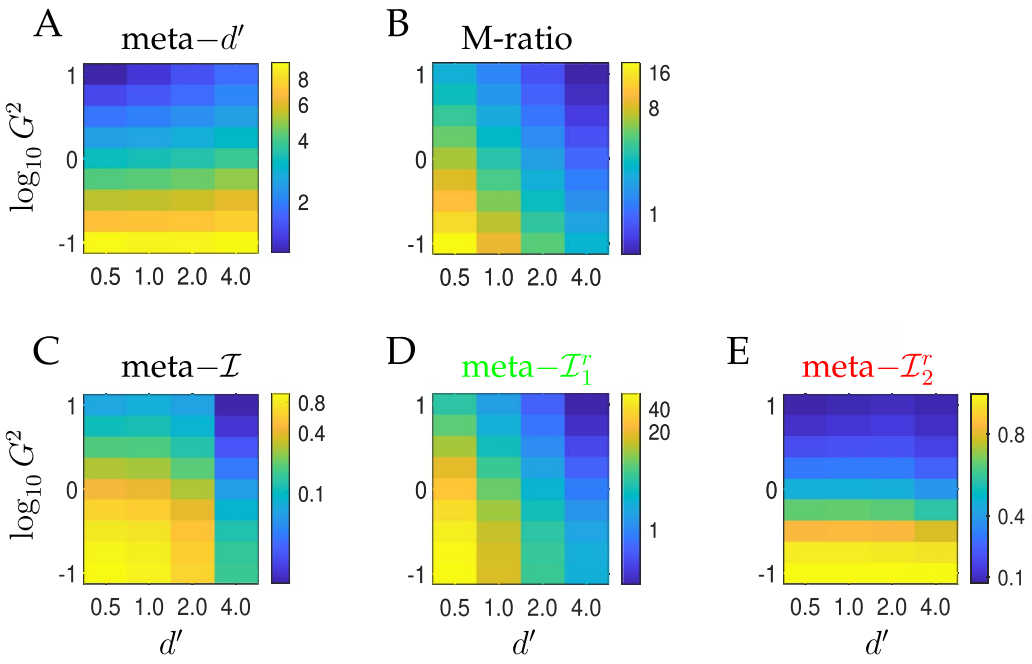


Figure 5. Meta-cognitive sensitivities and efficiencies for the second order observer as a function of d' and $\log_{10} G^2$ for the case that $B = 1$ and there are four confidence bins that are optimized to maximize meta–$\mathcal{I}$ (and so differ for each combination of d' and G^2). (A, B) meta–d' and the M-ratio, with the latter suggesting that the increase in meta–d' for larger d' is not a sign of efficiency. (C) meta–$\mathcal{I}$, showing that as d' gets large, the mutual information decreases, since the entropy of accuracy, $\mathcal{H}_2(r)$, decreases. (D) Normalizing meta–$\mathcal{I}$ by the mutual information meta–$\mathcal{I}(d')$ leads to values that are close to the M-ratio (B) away from the regime in which G^2 is small so the rater has access to higher quality information than the actor. (E) Normalizing meta–$\mathcal{I}$ by the entropy of the accuracy $\mathcal{H}_2(r)$ provides a measure of absolute efficiency which is roughly constant for small G^2, as the rater's unique information dominates.

As we saw in Figure 2F and 2G, one particular issue that spurs the use of meta–$\mathcal{I}$ is the effect of increasing the granularity of the confidence rating. Figure 6 examines this issue from the perspective of our second-order rater, showing how meta–$\mathcal{I}$ (the same would be true of the efficiency measures) increases with the number of confidence levels for different qualities of actor and different amounts of independent information provided to the rater (quantified as by G^2). Here, the thresholds defining the levels were again set to optimize meta–$\mathcal{I}$. In this idealized case, the extra levels are never harmful, but the degree to which they are helpful varies quite substantially. As we saw in Figure 2G, the increase is greater for lower d' (condition 1 in that figure); it also grows with G^2. Various factors are involved: for instance whether there is such certainty of the actor ($d' = 4$) or the rater ($G^2 \simeq 0.1$) that two levels suffice. Note that these

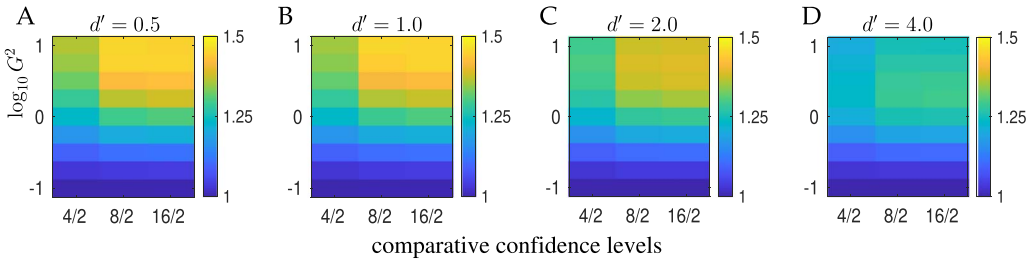


Figure 6. The effect of the number of levels of confidence. The ratio between meta–$\mathcal{I}$ (or equivalently meta–$\mathcal{I}^r_{\{1,2\}}$ values) for 4, 8, 16 confidence bins to that for 2 confidence bins for $d' = 0.5$ (A), $d' = 1$ (B) and $d' = 2$ (C) for 4, 8, or 16 levels of confidence (all optimized) and for values of G^2 between 0.1 and 10. The scales are set to be the same for all three heatmaps. Here $B = 1$.

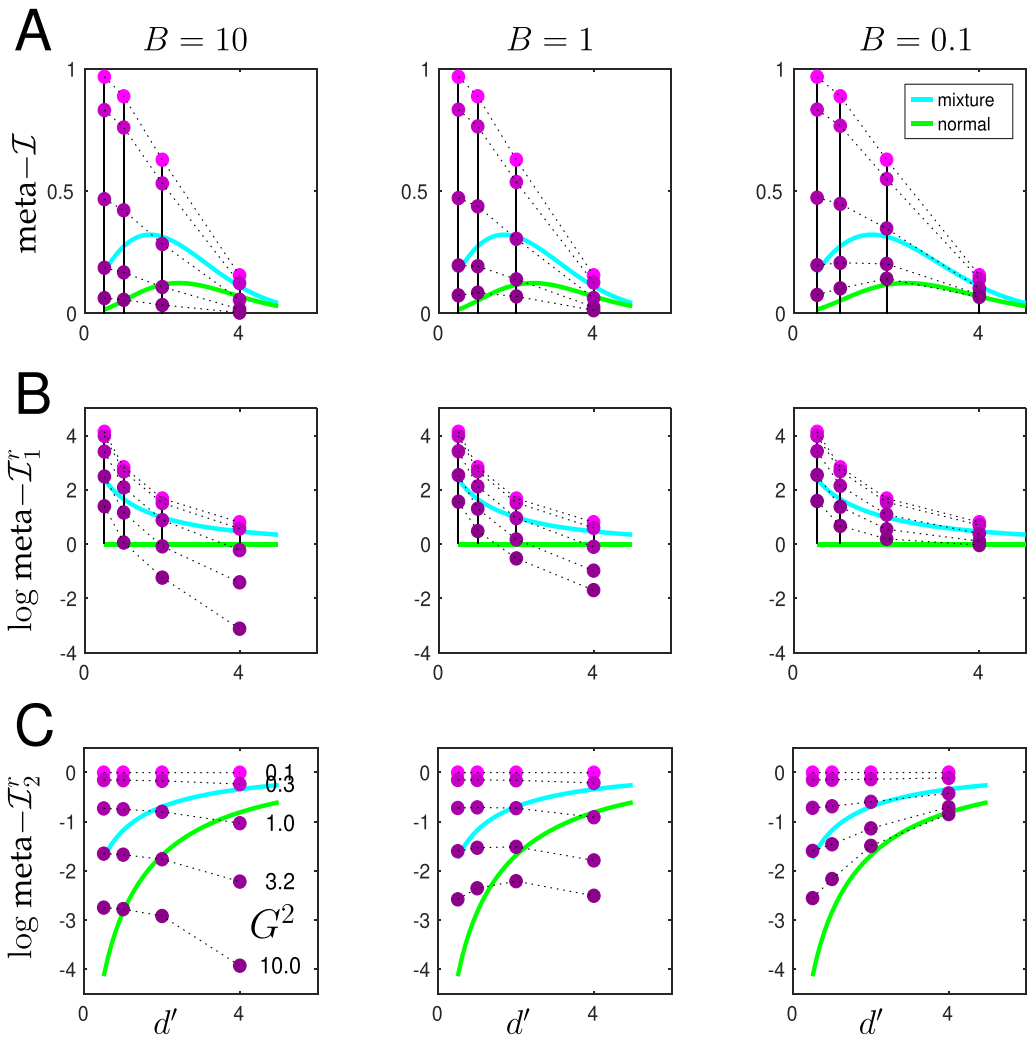


Figure 7. Hypo and hyper-sensitivity. The continuous lines show meta–$\mathcal{I}$ (A), log meta–$\mathcal{I}_1^r$ (B) and log meta–$\mathcal{I}_2^r$ (C) for the naive first order Bayesian case with continuous confidence levels across different values of d' for the case of standard signal detection theory (green; discriminating two Gaussian distributions with means $d = \pm 1$ and standard deviation $\sigma = 2/d'$) and for the extreme mixture case (cyan; with an actor that is 50% correct with probability $2(1 - \Phi(d'/2))$ and 100% correct otherwise). The magenta points are from the second order model with the three values of B shown in the titles and the values of $G^2 = \{10, 3.7, 1, 0.27, 0.1\}$, from bottom to (top) for the case of 16, optimally-spaced confidence levels. Points for the same value of G are connected by dotted lines for graphical convenience.

plots show the benefit as the ratios between 4, 8, 16 levels and 2 levels; as expected the total meta–$\mathcal{I}$ decreases as G^2 gets larger.

Even for a given number of levels, the mutual information can vary as a function of the actual confidence intervals. For instance, for the cases of Figures 3 and 4, if we use four evenly spaced levels ($c \in \{[0, 0.25), [0.25, 0.5), [0.5, 0.75), [0.75, 1.0]\}$) rather than the uneven ones ($c \in \{[0, 0.1), [0.1, 0.5), [0.5, 0.9)[0.9, 1.0]\}$) in the figures, meta–$\mathcal{I}$ increases to 0.12 bits for the case of Figure 3, and decreases to 0.60 bits for the case of Figure 4.

The efficiency measures meta–$\mathcal{I}_1^r$ and meta–$\mathcal{I}_2^r$ are ways of measuring the hypo- and hyper-metacognitive sensitivity for meta–$\mathcal{I}$. In Figure 7, we compare meta–$\mathcal{I}$ (A) and the efficiency ratios (B, C) for particular idealized actors and raters (green and cyan lines) with the same measures for the actual second order rater of the previous figures for the maximum number

(16) of optimized confidence bins (magenta points). The green line comes from the same rater that defined meta–$\mathcal{I}(d')$, i.e., the case of standard signal detection theory in which the actor discriminates two Gaussian distributions with means $d = \pm 1$ and common standard deviation $\sigma = 2/d'$, and the rater acts as a naive Bayesian type 1 rater with a continuous range of confidence levels. Thus, the green lines in Figure 7A are the same as in Figure 2D; and the green lines in Figure 7B are flat at 0, since meta–$\mathcal{I}^r_1$ is defined as the ratio between meta–$\mathcal{I}$ for a rater and meta–$\mathcal{I}(d')$ itself. The absolute efficiency of this idealized rater decreases as d' gets smaller, since the actor makes more errors, but the rater lacks the information to discriminate them.

The cyan lines show meta–$\mathcal{I}$ and meta–$\mathcal{I}^r_{\{1,2\}}$ for the less standard model of actor and type 1 rater that was considered as a *reductio ad absurdum* by Rahnev and Fleming (2019), for which the error rate associated with a conventional d' (which is $p_{\text{err}} = 1 - \Phi(d'/2)$) comes from a mixture model in which the actor and rater are either completely guessing (with actual and believed probability of error 50%); happening with a mixture proportion of $2p_{\text{err}}$; and actual and believed probability of error 0%, with a mixture proportion of $1 - 2p_{\text{err}}$. The values of meta–$\mathcal{I}$ and meta–$\mathcal{I}^r_{\{1,2\}}$ for this rater are higher than for the standard one, since there is extra information about the source of errors. This reminds us that d' is an incomplete measure of the actor's process, again making it important to interpret cautiously measures such as the M-ratio and meta–$\mathcal{I}^r_1$ that use it directly for normalization.

One could consider meta–$\mathcal{I}$ and meta–$\mathcal{I}^r_{\{1,2\}}$ values lower than these numbers to be hypo-efficient; and ones larger than these to be hyper-efficient—at least relative to these raters. The magenta points show that as G^2 decreases (bottom to top), the second order model generally goes from hypo to hyper-efficiency; but B also plays a role. This is clearest for the meta–$\mathcal{I}^r_1$, where for intermediate values of G, the rater becomes hypo-efficient as d' grows for large B, when the rater cannot prosper from the extra information the actor enjoys.

DISCUSSION

Fleming and Lau (2014) noted that standard ways of quantifying metacognition are based on distributions such as those in Figures 1 and 3F, 3H (at least if extra factors such as the time participants take to rate confidence are not taken into account; Desender et al., 2022). Indeed, Nelson (1984) listed eight such measures, which do not include meta–d' or meta–$\mathcal{I}$ or meta–$\mathcal{I}^r_{\{1,2\}}$, being measures of metacognitive sensitivity and efficiency that we advocate here. meta–$\mathcal{I}$ is the mutual information between the actual accuracy of the choices of the actor and the confidence ratings produced by the rater about those choices. This is a simple function of the same statistics used to calculate meta–d' and the M-ratio, and shares some of the desirable properties of those quantities (along, of course, with the less desirable ones rooted in assumptions such as that the decision-making process is stationary, with a fixed strength of evidence; Rahnev & Fleming, 2019). However, like correlation measures (Nelson, 1984), meta–$\mathcal{I}$ has the additional benefit of not depending on a potentially imperfect fit to a model of type 1 choice that might not be exactly appropriate, and it also scales appropriately with such factors as the number of levels of confidence. Note that Bowman et al. (2022) suggested using the mutual information to measure a form of type 1 sensitivity in a study of awareness rather than confidence.

That apparent metacognitive efficiency can increase with the number of levels with which subjects rate confidence suggests that in experiments collecting confidence ratings, it would be worth paying some extra attention to the way that reports are solicited. In the data from

Shekhar and Rahnev (2021), metacognitive efficiency increased up to 10 levels of reporting—at least on average—something we could observe clearly by looking at alternative quantizations of the nearly continuous confidence data they had collected. It would be interesting to carry out this exercise on other data. It should be noted that evaluating mutual information measures for a truly continuous confidence report is tricky from modest amounts of data, because of known biases (Kozachenko & Leonenko, 1987; Paninski, 2003, Panzeri & Treves, 1996; Witter & Houghton, 2021); and so further study in particular cases would be most valuable.

We illustrated meta–$\mathcal{I}$ and meta–$\mathcal{I}^r_{\{1,2\}}$, and compared them with the other measures, using both data from Shekhar and Rahnev (2021) and a simple case of a second order rater (Fleming & Daw, 2017), which is not restricted to having confidence at least as large as 50% (as would be true of a naive type 1 Bayesian), and can be either hypo- or hyper-metacognitively efficient.

The evaluation of the mutual information is completely bias-free. Of course, as noted in the Introduction, bias can affect the mutual information, by affecting the utilitization of the confidence levels, thereby increasing the conditional entropy of the accuracy given the rating ($\mathcal{H}(r|c, a)$). However, like all information theoretic quantities (though unlike the M-ratio; Xue et al., 2021), meta–$\mathcal{I}$ is completely unaffected by the *labels* that are given to the confidence levels—it is only influenced by the conditional accuracy that these levels afford. Indeed, meta–$\mathcal{I}$ would be unaffected if the labels were scrambled, so that subjects notionally reported 'high confidence' in the actor's choice to mean that an error was likely, and 'low confidence' when an error was not. This is also unlike meta–d', which makes assumptions about the appropriate monotonicity of the reporting levels in order to be able to calculate a notional type 1 d'. It might be possible to include in the optimization process that leads to the evaluation of meta–d' an additional reordering of the confidence levels—although this would then create a complex combinatorial optimization problem (with $n!$ possible orders for n levels). It would be interesting to examine other information theoretic mechanisms that might preserve at least the order of the levels. For instance, one might imagine the act of reporting as being like a noisy channel, in which subjects can stochastically report levels that are somewhat different from their true confidence. A treatment that would have this effect (albeit, technically, by varying the confidence criteria rather than the confidence signal) is exactly what was suggested by Shekhar and Rahnev (2021) as a process model for metacognitive inefficiency in the data that we analyzed above (see also Guggenmos, 2022). One limit of such noisy processes might offer a good formalization of the common empirical practice of recording confidence on a continuous scale (e.g., using a slider), but then for the experimenter to create a set of bins whose width would be determined by the structure of the stochastic report.

We noted that different positioning of the bins of confidence (even keeping the number constant) could lead to different values of meta–$\mathcal{I}$. We approximately optimized the bins on a case by case basis,[2] a bit like efficient coding of sensory information (Laughlin, 1981; Zhaoping, 2006), where coding levels are adjusted to reflect the 'natural' statistics of the information that is being coded. In the context of social inference, Bang et al. (2017) suggested maximising the entropy of confidence reports. The mutual information of Equation 1 can equivalently be written as the difference between the unconditional entropy of the confidence reports (given the action) and the conditional entropy given the accuracy (and the action). Thus maximizing the entropy can be beneficial for improving meta–$\mathcal{I}$. However, the consequences for the second term of the mutual information, namely the conditional entropy of

[2] Although in future work, it would also be worth using cross validation methods to estimate thresholds for discretizing confidence judgments.

the confidence reports given the accuracy (and the action), also need to be taken into account. It would certainly be interesting to examine the optimal meta–$\mathcal{I}$ solutions in more detail. If, as in Bang et al. (2017), the confidence regions change over time (for instance, to optimize their utilization), then metacognitive effectiveness and meta–$\mathcal{I}$ will change too, and would need to be tracked in an appropriately dynamic manner, something that poses potential statistical problems. It has been noted that the test-retest reliability of the M-ratio is compromised when the number of rating levels increases, suggesting that it would be important to investigate the equivalent for meta–$\mathcal{I}$ and meta–$\mathcal{I}^r_{\{1,2\}}$.

There is much discussion of the need for meta–d' to be corrected by the type 1 sensitivity d' in order to assess metacognitive efficiency—since cases with high d' are intrinsically easier. This has some undesirable consequences—for instance if d' is very low, then even a modest meta–d' can lead to an extremely large M-ratio, something that has inspired the use of the difference (meta–d' − d') rather than the M-ratio, or the logarithm of the M-ratio in such circumstances. We showed that meta–$\mathcal{I}^r_1$ has a similar issue and suggested that in the regime for the second-order model in which the rater is replete with its own sources of information that exceed those of the actor, it is more appropriate to consider the absolute efficiency of metacognition, normalizing meta–$\mathcal{I}$ instead by the unconditional entropy of the accuracy $\mathcal{H}_2(r)$, which is an upper bound to the mutual information, and is the total available variability that confidence could potentially rate. This alternative ratio meta–$\mathcal{I}^r_2$ assesses just how little of the available unconditional entropy of the accuracy is lost to a high conditional entropy ($\mathcal{H}_2(r|c, a = -1)$) in the mutual information equation.

One might look at both meta–$\mathcal{I}$ or the meta–$\mathcal{I}$-ratio as potential correlates of brain structure and function (Baird et al., 2015; Fleming & Dolan, 2012; Fleming et al., 2010). Note that the informational quantities can formally also accommodate tasks which use multiple d' values (for instance, if the quality of sensory information is different from trial to trial, as in the mixture curve of Figure 7). However, interpretative care is necessary (Rahnev & Fleming, 2019).

Like other information theoretic proposals in neuroscience, meta–$\mathcal{I}$ arguably offers more insight into bounds on the nature and quality of the computations involved in metacognition than into the neural realization of these computations. Process models such as those in Desender et al. (2022), Guggenmos (2022), and Shekhar and Rahnev (2021) or even the simple second order model that we considered (Fleming & Daw, 2017; Jang et al., 2012) are an attractive alternative, albeit adopting far stronger assumptions. Nevertheless, meta–$\mathcal{I}$ or meta–$\mathcal{I}^r_{\{1,2\}}$ would be drop-in replacements for other measures such as the M-ratio for such assessments as volume-based morphometry for regions whose size is correlated with the quality of metacognition (Fleming et al., 2010). Of course, despite the attractive theoretical properties of information theoretic measures, they are far from unique in measuring the quality of raters. Indeed, so-called skill scores (a term of art in assessing the sort of probabilistic forecasters with which metacognition is concerned) can be based on (strictly) proper scoring rules (Gneiting & Raftery, 2007) (a class including the famous quadratic Brier score; Brier (1950)); and the logarithmic scoring rule that underpins meta–$\mathcal{I}$). Correcting the evaluation of forecasters to reflect the difficulty of their forecasting tasks is also a concern in that literature.

In sum, the problem of confidence is inherently one of information—that the actor has about the true state of the stimulus; and that the rater has about the same quantity and about what the actor used. It therefore seems appropriate to use the methods of information theory to judge the relationship between the stimulus, the actor and the rater.

ACKNOWLEDGMENTS

I am very grateful to Dan Bang, Stephen Fleming, Liam Paninski and Lion Schulz for most helpful discussions and comments on earlier versions of the manuscript, and particularly to Matthias Guggenmos for an extensive discourse and him and Li Zhaoping for suggestions about normalization of meta–$\mathcal{I}$. Funding was from the Max Planck Society and the Humboldt Foundation. PD is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1—Project number 39072764 and of the Else Kröner Medical Scientist Kolleg "ClinbrAIn: Artificial Intelligence for Clinical Brain Research."

DATA AVAILABILITY STATEMENT

This study reused data that had been made publicly available based on the study of Shekhar and Rahnev (2021).

Data and code for reproducing the figures are available at https://github.com/Peter-Dayan-TN/metainf.git.

REFERENCES

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science*, *329*(5995), 1081–1085. https://doi.org/10.1126/science.1185718, PubMed: 20798320

Baird, B., Cieslak, M., Smallwood, J., Grafton, S. T., & Schooler, J. W. (2015). Regional white matter variation associated with domain-specific metacognitive accuracy. *Journal of Cognitive Neuroscience*, *27*(3), 440–452. https://doi.org/10.1162/jocn_a_00741, PubMed: 25313660

Bang, D., Aitchison, L., Moran, R., Herce Castanon, S., Rafiee, B., Mahmoodi, A., Lau, J. Y. F., Latham, P. E., Bahrami, B., & Summerfield, C. (2017). Confidence matching in group decision-making. *Nature Human Behaviour*, *1*(6), Article 0117. https://doi.org/10.1038/s41562-017-0117

Barrett, A. B., Dienes, Z., & Seth, A. K. (2013). Measures of metacognition on signal-detection theoretic models. *Psychological Methods*, *18*(4), 535–552. https://doi.org/10.1037/a0033268, PubMed: 24079931

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: An update. *Trends in Cognitive Sciences*, *8*(12), 539–546. https://doi.org/10.1016/j.tics.2004.10.003, PubMed: 15556023

Bowman, H., Jones, W., Pincham, H., Fleming, S., Cleeremans, A., & Smith, M. (2022). Modelling the simultaneous encoding/serial experience theory of the perceptual moment: A blink of meta-experience. *Neuroscience of Consciousness*, *2022*(1), Article niac003. https://doi.org/10.1093/nc/niac003, PubMed: 35242362

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Carter, C. S., Braver, T. S., Barch, D. M., Botvinick, M. M., Noll, D., & Cohen, J. D. (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science*, *280*(5364), 747–749. https://doi.org/10.1126/science.280.5364.747, PubMed: 9563953

Dawid, A. P. (1982). The well-calibrated Bayesian. *Journal of the American Statistical Association*, *77*(379), 605–610. https://doi.org/10.1080/01621459.1982.10477856

De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nature Neuroscience*, *16*(1), 105–110. https://doi.org/10.1038/nn.3279, PubMed: 23222911

Dehaene, S., Posner, M. I., & Tucker, D. M. (1994). Localization of a neural system for error detection and compensation. *Psychological Science*, *5*(5), 303–305. https://doi.org/10.1111/j.1467-9280.1994.tb00630.x

Desender, K., Donner, T. H., & Verguts, T. (2021). Dynamic expressions of confidence within an evidence accumulation framework. *Cognition*, *207*, Article 104522. https://doi.org/10.1016/j.cognition.2020.104522, PubMed: 33256974

Desender, K., Vermeylen, L., & Verguts, T. (2022). Dynamic influences on static measures of metacognition. *Nature Communications*, *13*(1), Article 4208. https://doi.org/10.1038/s41467-022-31727-0, PubMed: 35864100

Evans, S., & Azzopardi, P. (2007). Evaluation of a 'bias-free' measure of awareness. *Spatial Vision*, *20*(1–2), 61–77. https://doi.org/10.1163/156856807779369742, PubMed: 17357716

Fernandez-Duque, D., Baird, J. A., & Posner, M. I. (2000). Awareness and metacognition. *Consciousness and Cognition*, *9*(2), 324–326. https://doi.org/10.1006/ccog.2000.0449, PubMed: 10924252

Ferrell, W. R., & McGoey, P. J. (1980). A model of calibration for subjective probabilities. *Organizational Behavior and Human Performance*, *26*(1), 32–53. https://doi.org/10.1016/0030-5073(80)90045-8

Fleming, S. M. (2021). *Know thyself: The science of self-awareness*. Basic Books.

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review*, *124*(1), 91–114. https://doi.org/10.1037/rev0000045, PubMed: 28004960

Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, *367*(1594), 1338–1349. https://doi.org/10.1098/rstb.2011.0417, PubMed: 22492751

Fleming, S. M., Dolan, R. J., & Frith, C. D. (2012). Metacognition: Computation, biology and function. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, *367*(1594), 1280–1286. https://doi.org/10.1098/rstb.2012.0021, PubMed: 22492746

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, Article 443. https://doi.org/10.3389/fnhum.2014.00443, PubMed: 25076880

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science*, *329*(5998), 1541–1543. https://doi.org/10.1126/science.1191883, PubMed: 20847276

Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review*, *10*(4), 843–876. https://doi.org/10.3758/BF03196546, PubMed: 15000533

Gehring, W. J., Goss, B., Coles, M. G., Meyer, D. E., & Donchin, E. (1993). A neural system for error detection and compensation. *Psychological Science*, *4*(6), 385–390. https://doi.org/10.1111/j.1467-9280.1993.tb00586.x

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. https://doi.org/10.1198/016214506000001437

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. John Wiley.

Guggenmos, M. (2021). Measuring metacognitive performance: Type 1 performance dependence and test-retest reliability. *Neuroscience of Consciousness*, *2021*(1), Article niab040. https://doi.org/10.1093/nc/niab040, PubMed: 34858637

Guggenmos, M. (2022). Reverse engineering of metacognition. *eLife*, *11*, Article e75420. https://doi.org/10.7554/eLife.75420, PubMed: 36107147

Hoven, M., Lebreton, M., Engelmann, J. B., Denys, D., Luigjes, J., & van Holst, R. J. (2019). Abnormalities of confidence in psychiatry: An overview and future perspectives. *Translational Psychiatry*, *9*(1), Article 268. https://doi.org/10.1038/s41398-019-0602-7, PubMed: 31636252

Jang, Y., Wallsten, T. S., & Huber, D. E. (2012). A stochastic detection and retrieval model for the study of metacognition. *Psychological Review*, *119*(1), 186–200. https://doi.org/10.1037/a0025960, PubMed: 22059901

Kepecs, A., & Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, *367*(1594), 1322–1337. https://doi.org/10.1098/rstb.2012.0037, PubMed: 22492750

Kerns, J. G., Cohen, J. D., MacDonald III, A. W., Cho, R. Y., Stenger, V. A., & Carter, C. S. (2004). Anterior cingulate conflict monitoring and adjustments in control. *Science*, *303*(5660), 1023–1026. https://doi.org/10.1126/science.1089910, PubMed: 14963333

Kozachenko, L. F., & Leonenko, N. N. (1987). Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii*, *23*(2), 9–16.

Kunimoto, C., Miller, J., & Pashler, H. (2001). Confidence and accuracy of near-threshold discrimination responses. *Consciousness and Cognition*, *10*(3), 294–340. https://doi.org/10.1006/ccog.2000.0494, PubMed: 11697867

Lau, H. (2022). *In consciousness we trust: The cognitive neuroscience of subjective experience*. Oxford University Press. https://doi.org/10.1093/oso/9780198856771.001.0001

Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences*, *15*(8), 365–373. https://doi.org/10.1016/j.tics.2011.05.009, PubMed: 21737339

Laughlin, S. (1981). A simple coding procedure enhances a neuron's information capacity. *Zeitschrift für Naturforschung, Section C: Biosciences*, *36*(9–10), 910–912. https://doi.org/10.1515/znc-1981-9-1040, PubMed: 7303823

Mamassian, P., & de Gardelle, V. (2022). Modeling perceptual confidence and the confidence forced-choice paradigm. *Psychological Review*, *129*(5), 976–998. https://doi.org/10.1037/rev0000312, PubMed: 34323580

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422–430. https://doi.org/10.1016/j.concog.2011.09.021, PubMed: 22071269

Maniscalco, B., & Lau, H. (2014). Signal detection theory analysis of type 1 and type 2 data: Meta–d', response-specific meta–d', and the unequal variance SDT model. In S. M. Fleming & C. D. Frith (Eds.), *The cognitive neuroscience of metacognition* (pp. 25–66). Springer. https://doi.org/10.1007/978-3-642-45190-4_3

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, *95*(1), 109–133. https://doi.org/10.1037/0033-2909.95.1.109, PubMed: 6544431

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *Psychology of learning and motivation* (Vol. 26, pp. 125–173). Academic Press. https://doi.org/10.1016/S0079-7421(08)60053-5

Paninski, L. (2003). Estimation of entropy and mutual information. *Neural Computation*, *15*(6), 1191–1253. https://doi.org/10.1162/089976603321780272

Panzeri, S., & Treves, A. (1996). Analytical estimates of limited sampling biases in different information measures. *Network: Computation in Neural Systems*, *7*(1), 87–107. https://doi.org/10.1080/0954898X.1996.11978656, PubMed: 29480146

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: Distinct probabilistic quantities for different goals. *Nature Neuroscience*, *19*(3), 366–374. https://doi.org/10.1038/nn.4240, PubMed: 26906503

Rahnev, D., & Fleming, S. M. (2019). How experimental procedures influence estimates of metacognitive ability. *Neuroscience of Consciousness*, *2019*(1), Article niz009. https://doi.org/10.1093/nc/niz009, PubMed: 31198586

Rouault, M., Seow, T., Gillan, C. M., & Fleming, S. M. (2018). Psychiatric symptom dimensions are associated with dissociable shifts in metacognition but not task performance. *Biological Psychiatry*, *84*(6), 443–451. https://doi.org/10.1016/j.biopsych.2017.12.017, PubMed: 29458997

Schulz, L., Fleming, S. M., & Dayan, P. (2023). Metacognitive computations for information search: Confidence in control. *Psychological Review*, *130*(3), 604–639. https://doi.org/10.1037/rev0000401, PubMed: 36757948

Schulz, L., Rollwage, M., Dolan, R. J., & Fleming, S. M. (2020). Dogmatism manifests in lowered information search under uncertainty. *Proceedings of the National Academy of Sciences*, *117*(49), 31527–31534. https://doi.org/10.1073/pnas.2009641117, PubMed: 33214149

Seow, T. X. F., Rouault, M., Gillan, C. M., & Fleming, S. M. (2021). How local and global metacognition shape mental health. *Biological Psychiatry*, *90*(7), 436–446. https://doi.org/10.1016/j.biopsych.2021.05.013, PubMed: 34334187

Shekhar, M., & Rahnev, D. (2021). The nature of metacognitive inefficiency in perceptual decision making. *Psychological Review*, *128*(1), 45–70. https://doi.org/10.1037/rev0000249, PubMed: 32673034

Sun, X., Zhu, C., & So, S. H. W. (2017). Dysfunctional metacognition across psychopathologies: A meta-analytic review. *European*

Psychiatry, *45*, 139–153. https://doi.org/10.1016/j.eurpsy.2017.05.029, PubMed: 28763680

Vaccaro, A. G., & Fleming, S. M. (2018). Thinking about thinking: A coordinate-based meta-analysis of neuroimaging studies of metacognitive judgements. *Brain and Neuroscience Advances*, *2*, Article 2398212818810591. https://doi.org/10.1177/2398212818810591, PubMed: 30542659

Witter, J., & Houghton, C. (2021). A note on the unbiased estimation of mutual information. *arXiv:2105.08682*. https://doi.org/10.48550/arXiv.2105.08682

Xue, K., Shekhar, M., & Rahnev, D. (2021). Examining the robustness of the relationship between metacognitive efficiency and metacognitive bias. *Consciousness and Cognition*, *95*, Article 103196. https://doi.org/10.1016/j.concog.2021.103196, PubMed: 34481178

Zhaoping, L. (2006). Theoretical understanding of the early visual processes by data compression and data selection. *Network: Computation in Neural Systems*, *17*(4), 301–334. https://doi.org/10.1080/09548980600931995, PubMed: 17283516