Reverse engineering of metacognition

Matthias Guggenmos[1]

[1] Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Department of Psychiatry and Neurosciences, Charitéplatz 1, 10117 Berlin, Germany

Corresponding author:

Matthias Guggenmos

Charité Universitätsmedizin Berlin

Department of Psychiatry and Neurosciences

Address: Charitéplatz 1, 10117 Berlin, Germany

Phone: +49 30 450 517131

E-Mail: mg.corresponding@gmail.com

Abstract

The human ability to introspect on thoughts, perceptions or actions − metacognitive ability − has become a focal topic of both cognitive basic and clinical research. At the same time it has become increasingly clear that currently available quantitative tools are limited in their ability to make unconfounded inferences about metacognition. As a step forward, the present work introduces a comprehensive framework and model of metacognition that allows for inferences about metacognitive noise and metacognitive biases during the readout of type 1 decision values or at the confidence reporting stage. The model assumes that confidence results from a continuous but noisy and potentially biased transformation of decision values, described by a confidence link function. A canonical set of metacognitive noise distributions is introduced which differ, amongst others, in their predictions about metacognitive sign flips of type 1 decision values. Successful recovery of model parameters is demonstrated and the model is validated on an empirical data set. In particular, it is shown that metacognitive noise and bias parameters correlate with conventional behavioral measures. Crucially, in contrast to these conventional measures, metacognitive noise parameters inferred from the model are shown to be independent of type 1 performance. This work is accompanied by a toolbox (*ReMeta*) that allows researchers to estimate key parameters of metacognition in confidence datasets.

Introduction

The human ability to judge the quality of one's own choices, actions and percepts by means of confidence ratings has been subject to scientific inquiry since the dawn of empirical psychology (Pierce and Jastrow, 1885; Fullerton and Cattell, 1892), albeit it has long been limited to specific research niches. More recently, research on human confidence, and metacognition more generally, has accelerated and branched off to other domains such as mental illnesses (Rouault et al., 2018; Hoven et al., 2019; Moritz and Lysaker, 2019; Seow et al., 2021) and education (Fleur et al., 2021). Two main quantitative characteristics have emerged to describe subjective reports of confidence: *metacognitive bias* and *metacognitive sensitivity*.

Fullerton and Cattell (1892) already noted that "different individuals place very different meanings on the degree of confidence. Some observers are nearly always quite or fairly confident, while others are seldom confident." (p. 126). Technically, metacognitive biases describe a general propensity of observers towards lower or higher confidence ratings, holding the accuracy of the primary actions − type 1 performance − constant. From a perspective of statistical confidence, i.e. assuming that observers use confidence ratings to report probability correct, an observer is often considered *underconfident* or *overconfident* if confidence ratings are systematically below or above the objective proportion of correct responses.

Metacognitive biases of this type have been quite extensively studied in the judgement and decision making literature, in which they became known under the term *calibration* (Lichtenstein et al., 1977). A central finding is that humans have a tendency towards overestimating their probability of being correct (*overconfidence bias*), in particular in general knowledge questions (Lichtenstein et al., 1977, 1982; Harvey, 1997; but see Gigerenzer et al., 1991). More recently, overconfidence in decisions has been studied in psychiatric diseases, suggesting, for instance, underconfidence in individuals with depression (Fu et al., 2005, 2012; Fieker et al., 2016) and overconfidence in schizophrenic patients (Moritz and Woodward, 2006a; Köther et al., 2012; Moritz et al., 2014).

While the measurement of metacognitive biases is rarely problematized (more on this below), the intricacies of measuring metacognitive sensitivity have been the subject of critical discussion and have spurred a number of methodological developments (Nelson, 1984; Galvin et al., 2003; Maniscalco and Lau, 2012, 2014; Fleming and Lau, 2014). The issue is not the measurement of sensitivity per se: defining metacognitive (or type 2) sensitivity as the ability to discriminate between one's correct and incorrect responses, it is readily possible to compute this quantity using the logic of receiver operating curve analyses (type 2 ROC; Clarke et al., 1959; Pollack, 1959). The main issue is that metacognitive sensitivity, according to this definition, is strongly influenced by type 1 performance. The lower type 1 performance, the higher will be the number of guessing trials and thus the higher will also be the expected number of trials in which observers assign low confidence to accidental correct guesses. Expected metacognitive

sensitivity thus strongly depends on type 1 performance. Indeed, the importance of such type 1 performance confounds has been demonstrated in a recent meta-analysis of metacognitive performance aberrancies in schizophrenia (Rouy et al., 2020). The authors found that a previously claimed metacognitive deficit in schizophrenia was present only in studies that did not control for type 1 performance.

A potential solution to the problem of type 1 performance confounds was proposed by Maniscalco and colleagues through a measure called *meta-d'* (Rounis et al., 2010; Maniscalco and Lau, 2012, 2014). The idea of *meta-d'* is to express metacognitive performance in terms of the type 1 sensitivity that an ideal metacognitive observer would need in order to achieve the observed type 2 hit and false alarm rates. Since *meta-d'* is expressed in units of d', it can be directly compared to − and normalized by − type 1 sensitivity. The introduction of this new measure, which is used either as a ratio measure (M_{ratio} = *meta-d' / d'*) or as a difference measure (M_{diff} = *meta-d' / d'*), has since been widely used in the metacognition literature. Fleming and Lau (2014) coined the term referred *metacognitive efficiency* for such normalized measures.

Recently, however, these normalized measures have come under scrutiny. Bang et al. (2019) showed that the type 1 performance independence of M_{ratio} breaks down completely with the simple assumption of a source of metacognitive noise that is independent of sensory noise. Guggenmos (2021) confirmed in a recent systematic analysis of empirical (Confidence Database; Rahnev et al., 2020) and simulated data that M_{diff} is heavily dependent on type 1 performance. While M_{ratio} is more stable across different type 1 performance levels, it is not entirely independent either. The very same factor (metacognitive noise) that therefore plausibly introduces interindividual differences in metacognitive performance, might obviate a type-1-performance-independent measurement of metacognitive efficiency in this way. Apart from type 1 performance, a recent study has shown that in empirical data the *overall level of confidence* likewise affects M_{ratio} (Xue et al., 2021) − a confound that may be caused by different levels of metacognitive noise when overall confidence is low or high (Shekhar and Rahnev, 2021).

Here I argue that an unbiased estimation of latent metacognitive parameters requires a mechanistic forward model − a process model which specifies the transformation from stimulus input to the computations underlying confidence reports and which considers sources of metacognitive noise. In the current work, I introduce such a model. It allows researchers to make parametric inferences about metacognitive inefficiencies either during readout or during report, as well as about different types of metacognitive biases. The basic structure of the model is shown in Figure 1. It comprises two distinct levels for type 1 decision making (*sensory level*) and type 2 metacognitive judgments (*metacognitive level*).

A few key design choices deserve emphasis. First, the model assumes that confidence is a second-order process (Fleming and Daw, 2017) which assesses the evidence that guided type 1 behavior. In the proposed nomenclature of Maniscalco and Lau (2016) it corresponds to a hierarchical model and not to a

single-channel model in that it considers additional sources of metacognitive noise. A consequence of the hierarchical structure is that it is essential to capture the processes underlying the decision values at the type 1 level as precisely as possible, since decision values are the input to metacognitive computations. In the present model, this includes an estimate of both a sensory bias and a sensory threshold, both of which will influence type 1 decision values.

Second, recent work has demonstrated that metacognitive judgements are not only influenced by sensory noise, but also by metacognitive noise (Bang et al., 2019; Shekhar and Rahnev, 2021). In the present model, I therefore consider sources of metacognitive noise either during the readout of type 1 decision values or during report.

Third, human confidence ratings are often subject to metacognitive biases, which can lead to the diagnosis of underconfidence or overconfidence. I consider three different parameters that can be interpreted as metacognitive biases either at readout, at the stage of confidence computation or during report. The interpretation of these parameters as metacognitive biases entails the assumption that observers aim at reporting probability correct with their confidence ratings (*statistical confidence*; Hangya et al., 2016). Although I discuss link functions that deviate from this assumption, in the primary model outlined here, the transformation of sensory evidence to confidence therefore follows the logic of statistical confidence.

I demonstrate the issues of conventional measures of metacognitive ability and metacognitive bias, in particular their dependency on type 1 performance, and show that the process model approach can lead to unbiased inferences. Finally, I validate the model on a recently published empirical dataset (Shekhar and Rahnev, 2021). I illustrate for this dataset how model parameters can describe different facets of metacognition and assess the relationship of these parameters to conventional measures of metacognitive ability and metacognitive bias.

This article is accompanied by a toolbox − the Reverse engineering of Metacognition (*ReMeta*) toolbox, which allows researchers to apply the model to standard psychophysical datasets and make inferences about the parameters of the model. It is available at github.com/m-guggenmos/remeta.

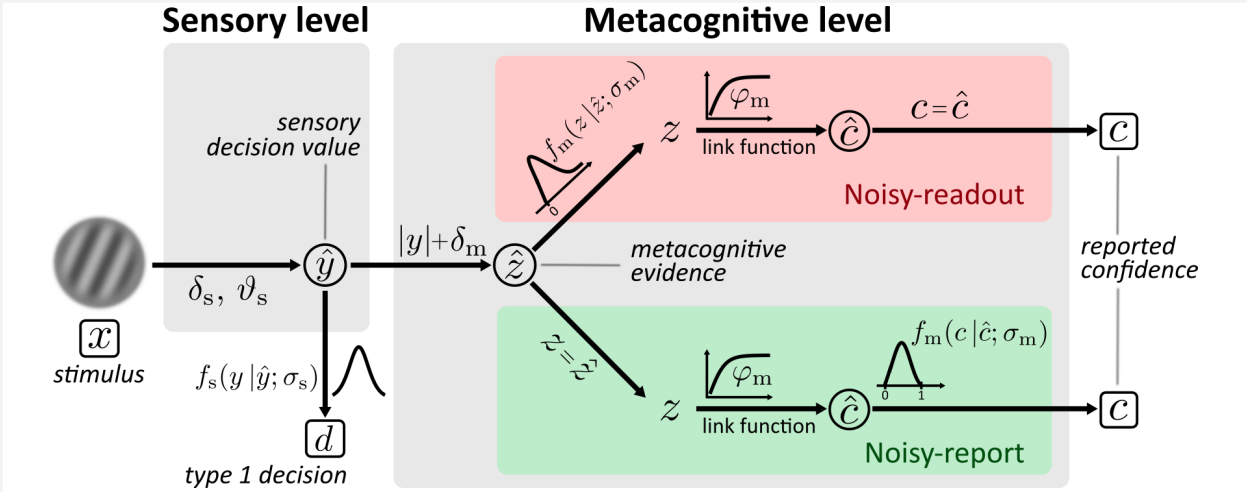

Figure 1. Computational model. Input to the model is the stimulus variable x, which codes the stimulus category (sign) and the intensity (absolute value). Type 1 decision-making is controlled by the sensory level. The processing of stimuli *x* at the sensory level is described by means of sensory noise (σ_s), bias (δ_s) and threshold (ϑ_s) parameters. The output of the sensory level is the decision value *y*, which determines type 1 decisions *d* and provides the input to the metacognitive level. At the metacognitive level it is assumed that the dominant source of metacognitive noise is either noise at the readout of decision values (*noisy-readout model*) or at the reporting stage (*noisy-report model*). In both cases, metacognitive judgements are based on the absolute decision value $|y|$, leading to an internal representation of metacognitive evidence $\hat{z}$. While the "readout" of this decision value is considered precise for the noisy-report model ($z = \hat{z}$), it is subject to metacognitive readout noise in the noisy-readout model ($z \sim f_m(\hat{z}; \sigma_m)$, where σ_m is a metacognitive noise parameter). Moreover, the readout might be subject to a first metacognitive bias described by the additive parameter δ_m. Finally, a link function – described by a multiplicative metacognitive bias φ_m – transforms metacognitive evidence to confidence $\hat{c}$. In the case of a noisy-report model, the dominant metacognitive noise source is during the report of confidence, i.e. confidence reports *c* are noisy expressions of the internal confidence representation $\hat{c}$.

Results

Results are structured in three parts. The first part introduces the architecture and the computational model, from stimulus input to type 1 and type 2 responses. The second part provides the mathematical basis for model inversion and parameter fitting and systematically assess the success of parameter recovery as a function of sample size and varied ground truth parameter values. Finally, in the third part, the model is validated on an empirical dataset (Shekhar and Rahnev, 2021).

1 Computational model

1.1 Computing decision values

For the model outlined here, the task space is restricted to two stimulus categories referred to as S^- and S^+. Stimuli are described by the stimulus variable x, the sign of which codes the stimulus category and the absolute value $|x|$ codes the intensity of the stimulus. The sensory level computes *decision values* $\hat{y}$ from the stimulus input x as follows:

$$\hat{y} = \begin{cases} x - \delta_s & \text{if } |x| > \vartheta_s \\ -\delta_s & \text{else} \end{cases} \quad (1)$$

The sensory bias parameter $\delta_s \in \mathbb{R}$ captures systematic preferences for one response category (Figures 2A,B,D). In addition, the sensory threshold $\vartheta_s \in \mathbb{R}^+$ defines the minimal stimulus intensity which is necessary to drive the system, i.e. above which the observer's type 1 choices can be better than chance level (Figures 2C,D). Note that a sensory threshold parameter should only be considered if the stimulus material includes intensity levels in a range at which participants perform close to chance. Otherwise the parameter cannot be estimated and should be omitted, i.e. Equation (1) reduces to $\hat{y} = x - \delta_s$.

In the model described here (henceforth referred to as the *primary model*) I assume that decision values can be linearly constructed from the stimulus variable x. In practise, this may often be too strong of an assumption and it may thus be necessary to allow for a nonlinear transformation of x ('nonlinear transduction', see e.g. Dosher and Lu, 1998). In supplementary section 1.1, I address this possibility with an additional nonlinear transformation parameter γ_s.

The final decision value y is subject to sources of sensory noise σ_s, described by a logistic distribution $f_s(y)$:

$$y \sim f_s(y) = \frac{\pi}{\sqrt{3}\sigma_s} \frac{\exp\left(\frac{\pi(y-\hat{y})}{\sqrt{3}\sigma_s}\right)}{\left(1+\exp\left(\frac{\pi(y-\hat{y})}{\sqrt{3}\sigma_s}\right)\right)^2} \quad (2)$$

Equation 2 is parameterized such that σ_s corresponds to the standard deviation of the logistic distribution. Figures 2E and 2F show two examples with low and high sensory noise σ_s, respectively. The logistic distribution was chosen over the more conventional normal distribution due to its explicit analytic solution of the cumulative density – the logistic function. In practise, both distributions are highly similar and which one is chosen is unlikely to matter.

In the primary model I assume that sensory noise σ_s is constant across the stimulus intensity spectrum. In supplementary section 1.2 I extend this notion to signal-dependent noise.

Type 1 decisions d between the stimulus categories S^+ and S^- are based on the sign of y:

$$d = \begin{cases} S^+ & \text{if } y \geq 0 \\ S^- & \text{if } y < 0 \end{cases} \tag{3}$$

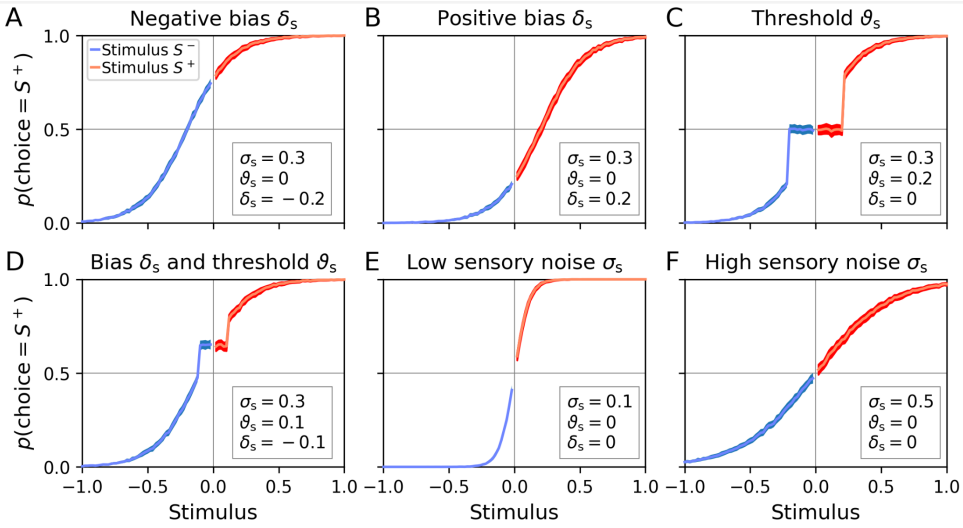

Figure 2. Simulated data exemplifying the effect of different model parameters of the sensory level on the probability of choosing S^+. Data are generated for 100 simulated observers and 10000 trials per observer. The probability of choosing S^+ is computed as the proportion of choices for S^+ for each simulated stimulus intensity. Shaded areas reflect the standard deviation across observers. Note that as an alternative to this simulation, choice probabilities can also be analytically computed by the posterior given in Equation 6. **(A)** A negative sensory bias δ_s shifts responses towards the stimulus category S^-. **(B)** A positive sensory bias δ_s shifts responses towards the stimulus category S^+. **(C)** Threshold ϑ_s. Stimulus intensities below the threshold lead to chance-level performance. **(D)** Bias δ_s and threshold ϑ_s. Example for simultaneous non-zero values of the bias and threshold parameter. **(E)** Model with a relatively low level of sensory noise σ_s. **(F)** Model with a relatively high level of sensory noise σ_s.

1.2 From decision values to metacognitive evidence

The decision values computed at the sensory level constitute the input to the metacognitive level. I assume that metacognition leverages the same sensory information that also guides type 1 decisions (or a noisy version thereof). Specifically, metacognitive judgements are based on a readout of absolute decision values $|y|$, respecing a metacognitive bias in the form of a readout term $\delta_m \in \mathbb{R}$:

$$\hat{z} = \max(|y| + \delta_m, 0) \tag{4}$$

Henceforth I refer to $\hat{z}$ as *metacognitive evidence*. Figure 3A illustrates the effect of the readout term δ_m on confidence.

1.3 The link function: from metacognitive evidence to confidence

The transformation from metacognitive evidence to predicted confidence $\hat{c}$ is described by a *link function*. A suitable link function must be bounded, reflecting the fact that confidence ratings always have lower and upper bounds, and increase monotonically.

I assume that observers aim at reporting probability correct, leading to a logistic link function in the case of the logistic sensory noise distribution (Equation 2). Without loss of generality, I use the range [0;1] for confidence ratings, such that a confidence of 0 indicates chance level probability (0.5) and a confidence of 1 the expectation of perfect type 1 performance. With these constraints and using the simple mathematical relationship between the logistic function and the tangens hyperbolicus, one arrives at the following link function:

$$\hat{c} = \tanh\left(\frac{\pi\varphi_{\mathrm{m}}}{2\sqrt{3}\sigma_{\mathrm{s}}}z\right) \tag{5}$$

Note that I use the variable z as opposed to $\hat{z}$, to indicate that the metacognitive evidence that enters the link function may be a noisy version of $\hat{z}$ (see the description of the *noisy-readout model* below). The *confidence slope* $\varphi_{\mathrm{m}} \in \mathbb{R}^{+}$ describes the steepness of the relationship between decision values and confidence (Figure 3B) and may be interpreted as a metacognitive bias (see Result section 1.5). For $\varphi_{\mathrm{m}} = 1$, Equation 5 describes confidence as computed by an ideal metacognitive observer.

Many other link functions are conceivable, which do not assume that observers aim at expressing confidence as probability correct. In particular, such link functions may not involve an estimate of sensory noise σ_{s}. Supplementary section 2 provides an overview of alternative link functions.

I refer to $\hat{c}$ as the *predicted confidence*, which may be different from the ultimately *reported confidence* c. This distinction becomes important when metacognitive noise is considered at the level of the report (see Result section 1.7).

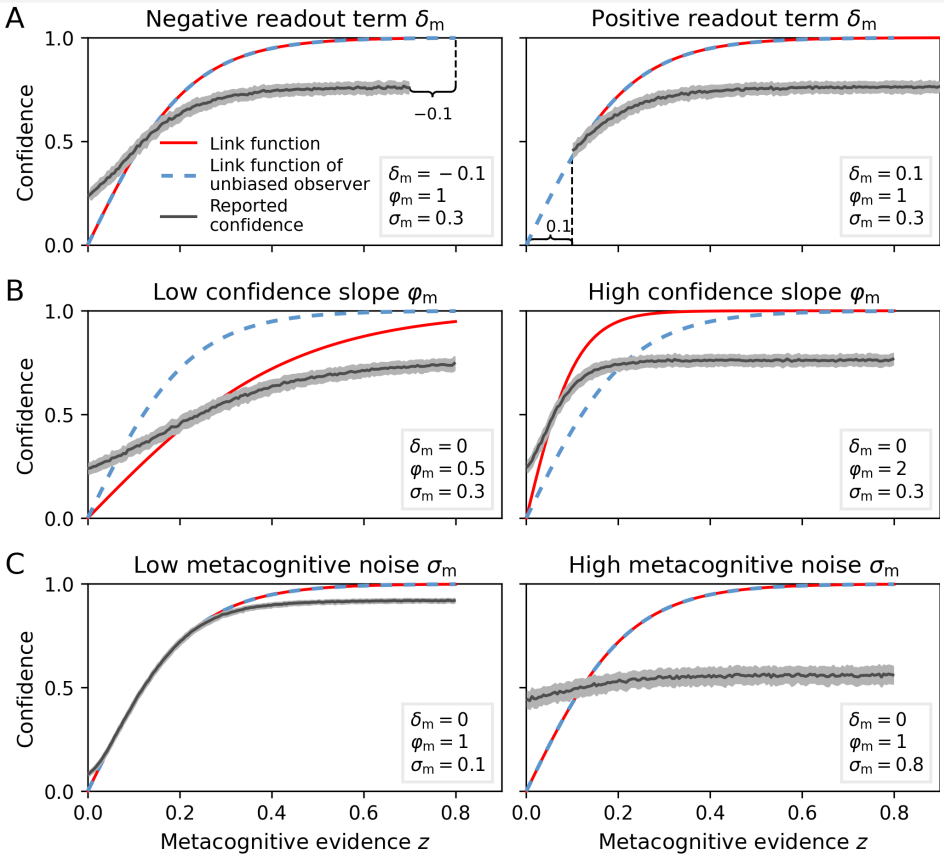

Figure 3. Metacognitive parameters and link function. Confidence data are generated for 100 simulated observers (10000 trials per observer) with noise added to confidence reports (see section 1.7). Shaded areas reflect the standard deviation across observers. Note that φ_m equals 1 for the link function of an ideal observer. **(A)** Readout term δ_m. The readout term is an additive term for the readout of decision values. Note that the readout term applies already at the level of decision values and hence does not affect the relationship between metacognitive evidence and confidence. Here, its effect is only visible in terms of a reduced maximum ($\delta_m < 0$) and increased minimum ($\delta_m > 0$) level of metacognitive evidence. **(B)** Confidence slope φ_m. A higher confidence slope leads to a steeper relationship between type 1 decision values and confidence. When assuming that an observer aims at computing a probability correct, the confidence slope may be interpreted as underconfidence ($\varphi_m < 1$) or overconfidence ($\varphi_m > 1$). **(C)** Metacognitive noise σ_m (here: noise at the stage of the report; cf. section 1.7). Higher metacognitive noise flattens the relationship between type 1 decision values and confidence. Note that the link function itself remains unaffected - the change in confidence is due to noise at the level of the confidence report, i.e. after the link function transformation.

1.4 Metacognitive biases

Metacognitive biases describe a systematic discrepancy between objective type 1 performance and subjective beliefs thereof (expressed via confidence ratings). Holding type 1 performance constant, overconfident observers systematically prefer higher confidence ratings, while underconfident observers systematically prefer lower confidence ratings. Importantly, metacognitive biases are orthogonal to the

metacognitive *sensitivity* of an observer. For instance, an underconfident observer who consistently chooses the second-lowest confidence rating for correct choices could have high metacognitive sensitivity nevertheless, as long as they consistently choose the lowest rating for incorrect choices.

Both δ_m and φ_m and in the model can be readily interpreted as a form of metacognitive bias. The readout term δ_m corresponds to an additive bias for the readout of decision values. As a result, metacognitive evidence is systematically increased or decreased by a constant δ_m. For $\delta_m < 0$, the readout term corresponds to a metacognitive threshold, i.e. the metacognitive level is only sensitive to absolute decision values above δ_m.

The confidence slope φ_m can either be regarded as a multiplicative bias for metacognitive evidence *z or* for the subject's belief about their own sensory noise σ_s. Applying the latter view, a value of $\varphi_m > 1$ would suggest that the observer underestimated sensory noise σ_s and hence shows overconfidence, whereas a value of $\varphi_m > 1$ implies that the observer overestimated σ_s and thus is underconfident.

In addition, one may consider a multiplicative bias parameter – a *confidence scaling parameter* λ_m – outside of the tangens hyperbolicus in Equation 5, i.e. $\lambda_m \cdot \tanh(..)$. As opposed to the confidence slope φ_m, which describes a scaling bias with respect to the internal estimate of sensory noise σ_s or metacognitive evidence *z*, λ_m describes a scaling bias of confidence at the stage of the report. As for the confidence slope φ_m, values of $\lambda_m < 1$ and $\lambda_m > 1$ reflect under- and overconfidence, respectively. The parameter λ_m is not part of the primary model, but it is part of the additional parameters specified in supplementary section 1.

To assess how the three metacognitive bias parameters relate to conventional measures of under- and overconfidence, I computed calibration curves (Lichtenstein et al., 1977) for a range of parameter values (Figure 4, left panels). A first observation concerns the case in which no metacognitive biases are present (i.e., $\delta_m = 0$, $\varphi_m = 1$, $\lambda_m = 1$; black lines). One could assume that calibration curves for bias-free observers are identical to the diagonal, such that objective and subjective accuracy are identical. This is not the case – the calibration curve is tilted towards overconfidence. This may seem surprising, but reflects exactly what is expected for a bias-free statistical confidence observer. This is best understood for the extreme case when the subjective probability correct is arbitrarily close to 1. Even for very high ratings of subjective probability there is – due to sensory noise – a certain finite probability that associated type 1 choices have been incorrect. Hence, objective type 1 performance is expected to be below the subjective probability in these cases. Importantly, *relative* to this bias-free observer all three metacognitive parameters yield calibration curves that resemble under- and overconfidence given appropriate choices of the parameter values (underconfidence: dotted lines; overconfidence: dashed lines).

As mentioned previously, metacognitive sensitivity (*AUROC2*, *meta-d'*) is strongly dependent on type 1 performance. How do metacognitive biases perform in this regard, when measured in a model-free manner from choice and confidence reports? To find out, I simulated confidence biases for a range of

metacognitive bias parameter values and type-1-performance levels (by varying the sensory noise parameter). Confidence biases were computed as the difference between subjective probability correct (by linearly transforming confidence from rating space [0; 1] to probability space [0.5; 1]) and objective probability correct. As shown in the middle panels of Figure 4, these results showcase the limits of naively measuring confidence biases in this way. Again, the bias-free observers shows an apparent overconfidence bias. In addition, this bias increases as type 1 performance decreases, reminiscent of the classic hard-easy effect for confidence (Lichtenstein and Fischhoff, 1977; for related analyses, see Soll, 1996; Merkle, 2009; Drugowitsch, 2016; Khalvati et al., 2021). At chance level performance, the overconfidence bias is exactly 0.25.

The value of 0.25 can be understood in the context of the '0.75 signature' (Hangya et al., 2016; Adler and Ma, 2018). When evidence discriminability is zero, an ideal Bayesian metacognitive observer will show an average confidence of 0.75 and thus an apparent (over)confidence bias of 0.25. Intuitively this can be understood from the fact that Bayesian confidence is defined as the area under a probability density in favor of the chosen option. Even in the case of zero evidence discriminability, this area will always be at least 0.5 – otherwise the other choice option would have been selected, but often higher.

The overconfidence bias leads to another peculiar case, namely that the bias of truly underconfident observers (i.e., $\delta_m < 0$, $\varphi_m < 1$, or $\lambda_m < 1$) can show a sign flip from over- to underconfidence as performance increases from chance level to perfect performance (dotted lines in the middle panels of Figure 4). Overall, the simulation underscores that metacognitive biases are just as confounded by type 1 behavior as metacognitive sensitivity.

Is it possible to recover unbiased estimates for the metacognitive bias parameters by inverting the process model? To find out, I again simulated data for a range of type-1-performance levels and true values of the bias parameters. In each case, I fitted the model to the data to obtain estimates of the parameters. As shown in the right panels of Figure 4, parameter recovery was indeed unbiased across the type 1 performance spectrum, with certain deviations only for extremely low or high type 1 performance levels. This demonstrates that, in principle, unbiased inferences about metacognitive biases are possible in a process model approach, under the assumption that the fitted model is a sufficient approximation of the empirical generative model.

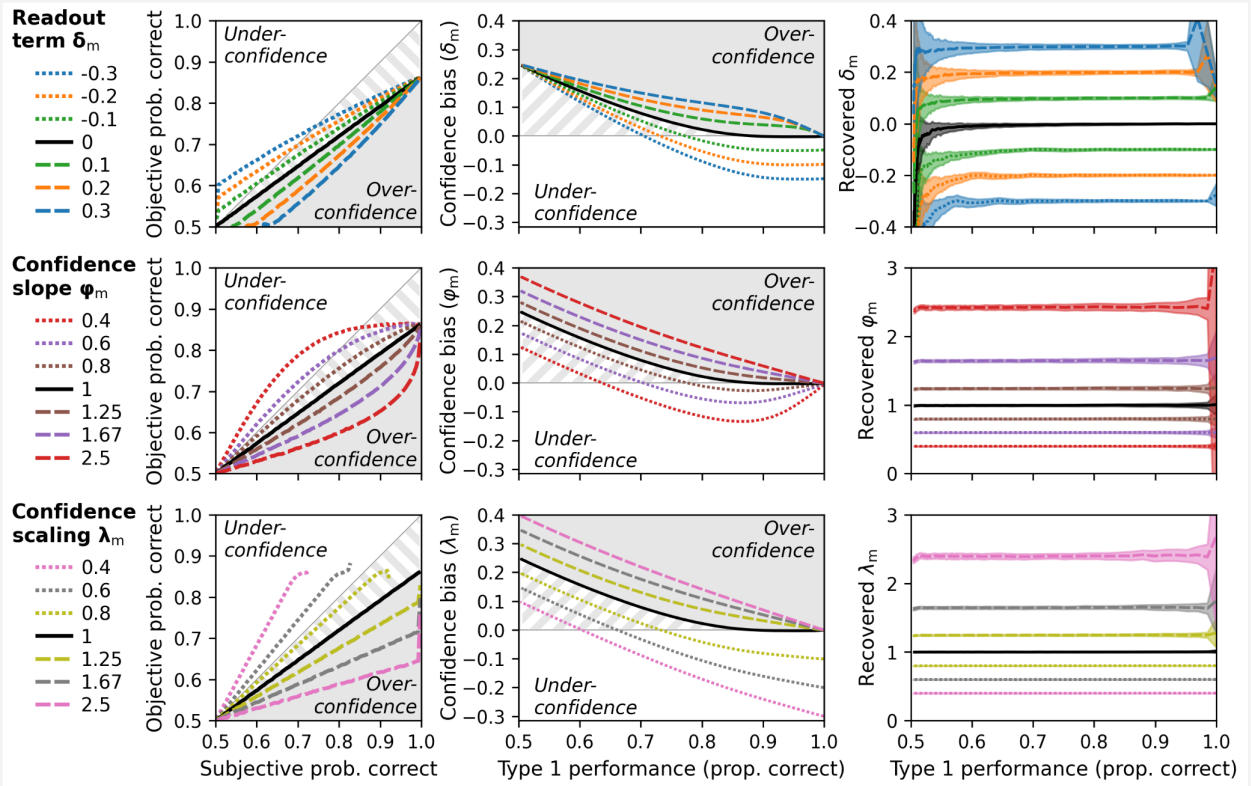

Figure 4. Metacognitive biases in the model. Grey shades indicate areas of overconfidence according to the model. Grey stripes areas indicate additional areas that would be classified as overconfidence in conventional analyses of confidence data. Simulations are based on a noisy-report + Beta distribution model. Metacognitive noise was set close to zero for simplicity. **(Left panels)** Calibration curves. Calibration curves compute the proportion of correct responses (objective probability correct) for each interval of subjective confidence reports. For this analysis, confidence was transformed from rating space [0; 1] to probability space [0.5; 1] and divided in 100 intervals with bin size 0.01. Calibration curves above and below the diagonal indicate under- and overconfident observers, respectively. Average type 1 performance for this simulation was around 70%. **(Middle panels)** Confidence bias in dependence of type 1 performance. Different levels of type 1 performance were simulated by sweeping the sensory noise parameter between 0.01 and 50. Confidence bias was computed as the difference between subjective probability correct and objective proportion correct. **(Right panels)** Recovery of metacognitive bias parameters in dependence of performance. Shades indicate standard deviations.

1.5 Confidence criteria

In the model outlined here, confidence results from a continuous transformation of metacognitive evidence, described by a parametric link function (Equation 5). The model thus has no confidence criteria. However, it would be readily possible to replace the tangens hyperbolicus with a stepwise link function where each step is described by the criterion placed along the z-axis and the respective confidence level (cf. Supplementary section 2; alternatively, one can assume equidistant confidence levels, thereby saving half of the parameters). Such a link function might be particularly relevant for

discrete confidence rating scales where participants associate available confidence ratings with often idiosyncratic and not easily parameterizable levels of metacognitive evidence.

However, even for the parametric link function of a statistical confidence observer it is worth considering two special confidence criteria: a minimum confidence criterion, below which confidence is 0, and a maximum criterion, above which confidence is 1. Indeed, the overproportional presence of the most extreme confidence ratings that is often observed in confidence datasets (cf. Confidence Database; Rahnev et al., 2020) may motivate such criteria.

My premise here is that these two specific criteria can be described as an implicit result of metacognitive biases. In general, when considering an ideal statistical confidence observer and assuming continuous confidence ratings, the presence of any criterion reflects suboptimal metacognitive behavior – including a minimum or maximum confidence criterion. According to Equation 5, an ideal observer's confidence should never be exactly 1 (for finite sensory noise) and should only ever be 0 when metacognitive evidence is exactly zero, which makes a dedicated criterion for this case likewise superfluous.

Importantly, a minimum confidence criterion is implicit to the readout term δ_m. As explained above, a negative value of δ_m effectively corresponds to a metacognitive threshold, such that metacognitive evidence z (and hence confidence) is zero for decision values smaller than δ_m. While a maximum confidence criterion is not covered by the primary model, it can be realized by the optional confidence scaling parameter λ_m considered in the previous section. Specifically, assuming $\lambda_m > 1$, the maximum criterion is the point along the metacognitive evidence axis at which a link function of the form $\lambda_m \cdot \tanh(..)$ becomes exactly 1. In sum, both a minimum and a maximum confidence criterion can be implemented as a form of a metacognitive bias.

1.6 Metacognitive noise: noisy-readout models

A key aspect of the current model is that the transformation from sensory decision values to confidence reports is subject to sources of metacognitive noise. In this section, I first consider a model of type *noisy-readout*, according to which the metacognitive noise mainly applies to the metacognitive readout of absolute sensory decision values (i.e., $\hat{z}$). The final metacognitive evidence z is thus a noisy version of $\hat{z}$. By contrast, the additional noise of reporting the confidence is considered negligible and the predicted confidence estimate $\hat{c}$ resulting from the link function is equal to the reported confidence c.

Metacognitive noise is defined by a probability distribution and a metacognitive noise parameter σ_m. The appropriate noise distribution for such readout noise is an open empirical question. Here, I introduce a family of potential candidates. A key consideration for the choice of a noise distribution is the issue of sign flips. Specifically, one could assume the metacognitive level initially deals with signed decision values, such that metacognitive noise can cause sign flips of these decision values. For instance, while an observer may have issued a type 1 response for stimulus category S^+, readout noise could flip the sign of

the decision value towards S^- at the metacognitive level. This scenario is akin to the case of error monitoring in the sense that an observer – by virtue of their metacognitive representation – will think to have made an error in their type 1 response (note that this is to be distinguished from post-decisional evidence accumulation, in which sign flips are based on *new* information, not noise).

How would an observer indicate their confidence in the case of a metacognitive sign flip? Unless confidence rating scales include the possibility to indicate errors (which I do not consider here), the only sensible response would be to indicate a confidence of 0, since confidence ratings apply to the choice made and not to the choice one would have hypothetically made based on a subsequent metacognitive representation. Overall, I thus take the following stance. Either one accepts the possibility of sign flips, but then only considers metacognitive evidence for the chosen option, which would be zero in the case of a sign flip; or, one rejects the possibility of sign flips altogether and assumes that the nature of metacognitive readout noise makes sign flips impossible. I assume that one of the two possibilities is true, thereby justifying the decision to only consider absolute decision values $\hat{z} = |y|$ for the chosen option at the metacognitive level.

If one accepts the possibility of sign flips, but assumes metacognitive evidence for the chosen stimulus to be zero in these instances, one arrives at the concept of a *censored* (or *rectified*) distribution. If a distribution is (left-)censored at zero, all negative parts of the distribution are assigned to the probability mass of zero, resulting in a distribution with a discrete term (at $z = 0$) and a continuous term ($z > 0$) (Figure 5A). In case of a normal distribution, the probability of z being exactly zero is equal to the cumulative density of the normal distribution at zero. An alternative to the normal distribution is a double exponential distribution, which allows for tail asymmetry. In particular, I here consider the Gumbel distribution which has a pronounced right tail, a property that fits recent observations regarding the skewed nature of metacognitive noise (Shekhar and Rahnev, 2021; Xue et al., 2021). Mathematical definitions of all distributions are listed in supplementary Table S2.

As outlined above, the second possibility is that the nature of metacognitive readout noise makes sign flips impossible, sparing the necessity of censoring. This required noise distributions that are bounded at zero, either naturally or by means of truncation. I first consider truncated distributions, in particular the truncated normal and the truncated Gumbel distribution (Figure 5B). Truncating a distribution means to cut off the parts of the distribution outside the truncation points (here the range below zero) and to renormalize the remainder of the distribution to 1.

While truncated distributions behave well mathematically, compared to censored distributions it is much less clear how a natural process could lead to a truncated metacognitive noise distribution. Truncated distributions occur when values outside of the bounds are discarded, which clearly does not apply to confidence ratings. I thus consider truncated distributions as an auxiliary construct at this point that may nevertheless qualify as an approximation to an unknown natural process.

Finally, there are many candidates of probability distributions that are naturally bounded at zero, perhaps the most prominent one being the lognormal distribution. In addition, I consider the Gamma distribution (Figure 5C), which has a more pronounced lower tail and is also the connatural counterpart to the Beta distribution for noisy-report models (see next section).

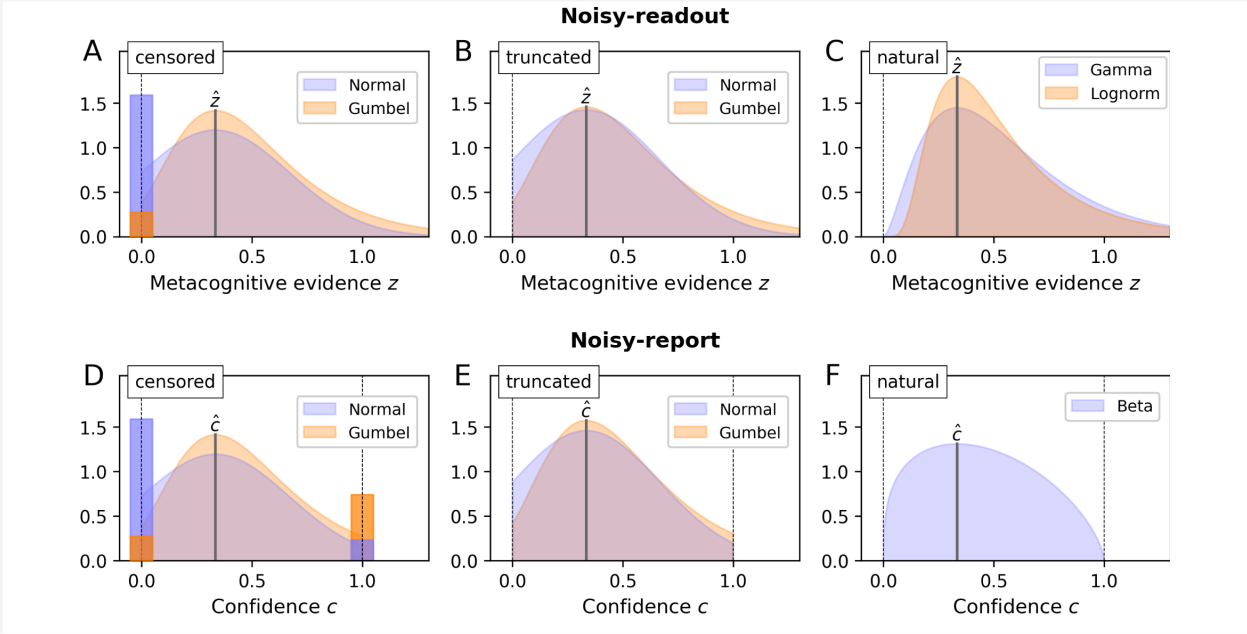

Figure 5. Metacognitive noise. Considered noise distributions are either censored, truncated or naturally bounded. In case of censoring, protruding probability mass accumulates at the bounds (depicted as bars with a darker shade; the width of these bars was chosen such that the area corresponds to the probability mass). The parameter σ_m and the distributional mode was set to ⅓ in all cases. **(A - C)** Noisy-readout models. Metacognitive noise is considered at the level of readout, affecting metacognitive evidence z. Only a lower bound at $z = 0$ applies. **(D - F)** Noisy-report models. Metacognitive noise is considered at the level of the confidence report, affecting internal confidence representations c. Confidence reports are bounded between 0 and 1.

1.7 Metacognitive noise: noisy-report models

In contrast to noisy-readout models, a noisy-report model assumes that the readout noise of decision values is negligible ($z = \hat{z}$) and that the dominant source of metacognitive noise occurs at the reporting stage ($c \sim f_m(\hat{c})$). Reporting noise itself may comprise various different sources of noise, occurring e.g. during the mental translation to an experimental confidence scale or in the form of visuomotor noise (e.g. when using a mouse cursor to indicate a continuous confidence rating).

A hard constraint for reporting noise is the fact that confidence scales are typically bounded between a minimum and a maximum confidence rating (reflecting the bounds [0; 1] for c in the present model). Reported confidence cannot be outside these bounds, regardless of the magnitude of reporting noise. As in the case of the noisy-readout model, one may consider either censored (Figure 5D), truncated

(Figure 5E) or naturally bounded distributions (Beta distribution; Figure 5F) to accommodate this constraint.

1.8 Metacognitive noise as a measure of metacognitive ability

As outlined above, I assume that metacognitive noise can be described either as variability during readout or report. In both cases, metacognitive noise is governed by the parameter σ_m. Higher values of σ_m will lead to a flatter relationship between reported confidence and decision values, i.e. confidence ratings become more indifferent with regard to different levels of internal sensory evidence (Figure 3B).

The behavior of the metacognitive noise parameter is closely related to the concept of metacognitive efficiency (Fleming and Lau, 2014), a term coined for two measures of metacognitive ability that aim at being invariant to type 1 performance: M_{ratio} and M_{diff}. As outlined in the introduction, the independence of type 1 performance has been contested for both measures, on the basis of empirical data and as well as in simulations that consider the presence of metacognitive noise (Bang et al., 2019; Guggenmos, 2021).

Here, I was interested in two main questions: can metacognitive noise σ_m be truthfully recovered regardless of type 1 performance? And further, to what degree are metacognitive noise σ_m and metacognitive efficiency correlated and thus potentially capture similar constructs?

To assess the type 1 performance dependency, I simulated data with varying levels of sensory noise σ_s and five different values of σ_m. In each case I computed M_{ratio} on the data and also fitted the model to recover metacognitive noise[1]. As shown in Figures 6A (noisy-report) and S3A (noisy-readout), one can reproduce the type 1 performance dependency of M_{ratio}, which is characterized by an initial increase and a subsequent decrease of M_{ratio} as type 1 performance increases (Guggenmos, 2021).

By contrast, as shown in Figure 6B for the noisy-report model, the parameter σ_m is recovered without bias across a broad range of type 1 performance levels and at different levels of generative metacognitive noise. For the noisy-readout model, recovery likewise works also well across a large range of type 1 performance levels; however, for very low or very high type 1 performance, recovery becomes biased and unstable (Supplementary Figure S5B). A likely reason is related to the inversion of the link function, which is necessary for parameter inference in noisy-readout models: since the link function is dependent on sensory noise σ_s, its inversion becomes increasingly ambiguous as σ_s becomes either very small or very large. However, apart from these extremal cases under the noisy-readout model, σ_m is largely unbiased and is thus a promising candidate to measure metacognitive ability independent of type 1 performance.

[1] Note that I focus on M_{ratio}, as the type 1 performance dependency is even more severe for M_{diff} (Guggenmos, 2021).

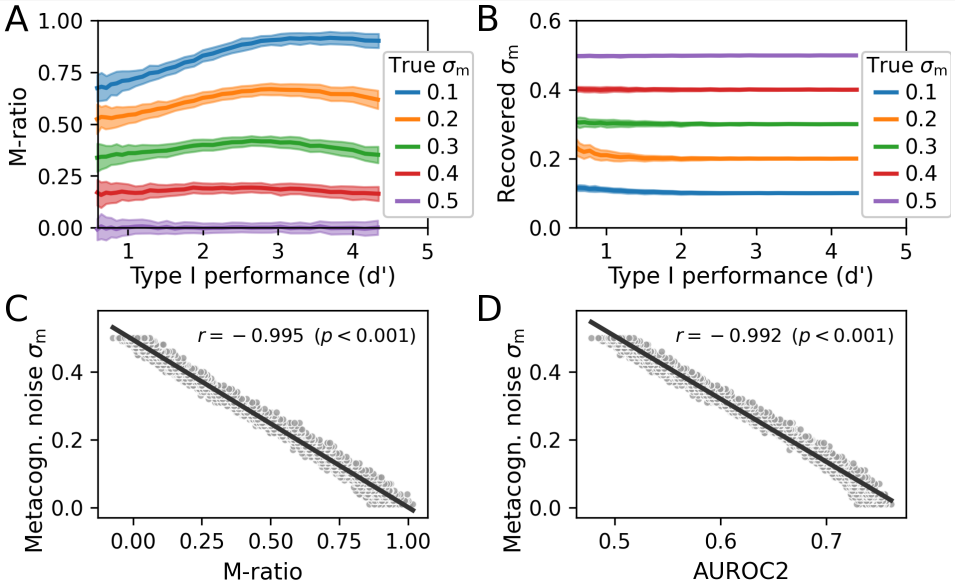

Figure 6. Metacognitive noise σ_m (noisy-report model) in comparison to conventional measures of metacognitive performance. Simulations for this figure were performed with a noisy-report model and a Beta distribution for metacognitive noise. Differences to other noise distributions are negligible. **(A, B)** Type 1 performance dependency. Different performance levels were induced by varying the sensory noise of the forward model. Five different levels of metacognitive noise were tested, ranging from relatively little metacognitive noise (σ_m= 0.1) to relatively high metacognitive noise (σ_m= 0.5). While M_{ratio} decreased with increasing type 1 performance (panel A), the recovered metacognitive noise parameter σ_m was largely independent of type 1 performance (panel B). Shaded areas indicate standard deviations across 100 simulated subjects. **(C, D)** Relationship between metacognitive noise and conventional measures of metacognitive performance. Simulated data were generated with a range of varying metacognitive noise parameters σ_m (0 to 0.5 in steps of 0.01) and constant sensory noise (proportion correct responses: 0.82). Based on these data, *AUROC2* values and M_{ratio} values were computed as measures of metacognitive sensitivity and efficiency, respectively. There is a clear negative correspondence between σ_m and both measures of metacognitive performance, reflecting the fact that metacognitive performance decreases with higher metacognitive noise.

Despite the fact that M_{ratio} may not be entirely independent of type 1 performance, it is fairly likely that it will capture the metacognitive ability of observers *to some degree.* It is thus interesting to assess the relationship between the model-based measure of metacognitive noise (σ_m) and M_{ratio}. To this aim, I performed a second simulation which kept type 1 performance constant at around 82% correct by using a fixed sensory noise parameter ($\sigma_s = 0.5$) and only varied the true underlying σ_m across a broad range between $\sigma_m = 0.01$ and $\sigma_m = 0.5$ in steps of 0.01. In addition, M_{ratio} was computed for each simulated observer. As shown in Figures 6C and 6D, there was indeed a strong negative correlation between σ_m and M_{ratio} (r = −0.983), as well as between σ_m and *AUROC2* (r = −0.979). The negative sign of the correlation is expected since a higher degree of noise should lead to more imprecise confidence ratings and thus reduced metacognitive performance.

2 Parameter fitting

Parameter fitting proceeds in a two-stage process. First, parameters of the sensory level are fitted by maximizing the likelihood of the model with respect to the observed type 1 decisions. Second, using the decision values predicted by the sensory level, the parameters of the metacognitive level are fitted by maximizing the likelihood with respect to observed confidence reports. The two levels are thus fitted independently.

In the following, the capital letter *D* denotes observed type 1 decisions and the capital letter *C* denotes observed confidence ratings. The set of parameters of the sensory level is denoted as $\mathcal{P}_s$:={σ_s, δ_s, ϑ_s} and the set of parameters of the metacognitive level as $\mathcal{P}_m$:={σ_m, δ_m, φ_m}.

2.1 Sensory level

At the sensory level, sensory noise is considered to follow a logistic distribution (Equation 2). The likelihood $\mathcal{L}$ of a particular type 1 decision D for stimulus x has an analytic solution given by the logistic function:

$$\mathcal{L}\left(D = S^{+} \mid \mathcal{P}_{\mathrm{s}}\right) = 1 - \mathcal{L}\left(D = S^{-} \mid \mathcal{P}_{\mathrm{s}}\right) = \frac{1}{1 + \exp\left(-\frac{\pi}{\sqrt{3}\sigma_{\mathrm{s}}}\hat{y}(x; \vartheta_{\mathrm{s}}, \delta_{\mathrm{s}})\right)} \qquad (6)$$

where $\hat{y}(x; \vartheta_s, \delta_s)$ is given by Equation 1. By maximizing the (cumulative) likelihood, one obtains estimates for ϑ_s, δ_s and σ_s.

2.2 Metacognitive level

Parameter inference at the metacognitive level requires the output of the sensory level (decision values *y*), an estimate of sensory noise σ_s and empirical confidence ratings *C*. By running the model in feed-forward mode and using the fitted sensory parameters, the likelihood of confidence ratings is evaluated either at the stage of readout (noisy-readout model) or report (noisy-report model).

Special consideration is necessary for the noisy-readout model in which the significant metacognitive noise source is assumed at the level of an unobserved variable − metacognitive evidence. For this reason, the model has to be inverted from the point of the observed variable (here confidence ratings) into the space of the latent variable (metacognitive evidence). A consequence of this is that the link function that transforms metacognitive decision values to confidence ratings must be strictly monotonically increasing in the noisy-readout scenario, as model inversion would otherwise be ambiguous.

Using the link function considered for this work, the tangens hyperbolicus (Equation 5), the inversion is as follows:

$$Z = \frac{2\sqrt{3}\sigma_{\mathrm{s}}}{\pi\varphi_{\mathrm{m}}}\operatorname{arctanh}(C) \qquad (7)$$

The condition of strict monotonicity is also the reason why the readout term δ_m is not considered a parameter of the link function: since confidence has a lower bound of 0, a hypothetical function of the form $\hat{c} = \tanh(z - \delta_m)$ would have to be rectified to 0 for all $z < \delta_m$ and thus would be non-invertible.

Importantly, the likelihood $\mathscr{L}(C \mid \mathcal{P}_m)$ of observed confidence ratings C given parameters $\mathcal{P}_m$ not only depends on the uncertainty of the model prediction for metacognitive decision values $\hat{z}(y)$, but also on the uncertainty of decision values y. Computing the likelihood $\mathscr{L}(C \mid \mathcal{P}_m)$ thus requires an integration over the probability density $f_s(y)$ of all valid decision values y:

$$\text{Noisy-readout:} \qquad \mathcal{L}\big(C \mid \mathcal{P}_\mathrm{m}\big) = \int_{\text{valid y}} f_\mathrm{m}\big(Z \mid \hat{z}(y)\big) f_\mathrm{s}(y)\,\mathrm{d}y \tag{8}$$

Valid decision values are those congruent with the empirical choice D.

In case of the noisy-report model, the likelihood can be directly computed with respect to the observed variable C, i.e. without inversion of the link function:

$$\text{Noisy-report:} \qquad \mathcal{L}\big(C \mid \mathcal{P}_\mathrm{m}\big) = \int_{\text{valid y}} f_\mathrm{m}\big(C \mid \hat{c}(y)\big) f_\mathrm{s}(y)\,\mathrm{d}y \tag{9}$$

2.3 Parameter recovery

Parameter recovery is an important part of model validation and refers to the process of generating data from the model with known parameters and then recovering those parameters through model fitting. Here, I tested parameter recovery by generating data with the forward model across a range of sample sizes and true parameter values.

In a first analysis, each of the six considered parameters of the primary model (σ_s, ϑ_s, δ_s, σ_m, δ_m, φ_m) was set to 0.2 in the generative model and parameter recovery was tested for a range of sample sizes between 500 and 10,000 trials. As shown in Supplementary Figure S6, parameter recovery was unbiased such that the parameter average across 1000 iterations of this procedure was always close to 0.2. Second, as one would expect, the precision of parameter reconstruction increased with sample size.

Since a generic parameter value of 0.2 is necessarily arbitrary, in a second analysis the sample size was fixed to 10,000 and the generative parameter values were systematically varied. As shown in Supplementary Figure S7, this analysis confirmed that parameter recovery was accurate across a sensible range for all parameters.

3 Application to an empirical dataset

To test the proposed model on real-world empirical data, I used a data set recently published by Shekhar and Rahnev (2021) which has a number of advantageous properties for a modeling approach. First, a high number of 2800 trials was measured for each of the 20 participants, enabling a precise estimate of computational parameters (cf. Supplementary Figure S6). Second, the task design comprised multiple

stimulus intensities, which is an important prerequisite for process models. And third, participants rated their confidence on a continuous scale. While the model works well with discrete confidence ratings, only continuous confidence scales harness the full expressive power of the model. In each trial, participants indicated whether a Gabor patch imposed on a noisy background was tilted counterclockwise or clockwise from a vertical reference and simultaneously rated their confidence on a continuous scale. The average performance was 77.7% correct responses.

Figure 7A visualizes the overall model fit at the sensory level. The posterior, defined as the probability of choosing S^+, closely matches the model fit. The average posterior probability shows a slight y-offset, reflecting a negative mean sensory bias towards S^+ ($\delta_s = -0.06 \pm 0.03$) and serving as a validation that the model parameter sensibly captured the psychometric curve. Since no stimulus intensities near chance-level performance were presented to participants, a sensory threshold parameter was not fitted.

At the metacognitive level, I compared noisy-readout and noisy-report models in combination with the metacognitive noise distributions introduced in Result sections 1.6 and 1.7. The model evidence was computed based on the Akaike information criterion (AIC; Akaike, 1974). As shown in Figure 7B, with the exception of censored distributions, all models performed at a similar level. Seven of the 10 tested models were the winning model for at least one participant (Figure 7C).

Interestingly, there were quite clear patterns between the shapes of individual confidence distributions and the respective winning model (Supplementary Figure S8). For instance, a single participant was best described by a noisy-report+Beta model, and indeed the confidence distribution of this participant is quite unique and plausibly could be generated by a Beta noise distribution (participant 7). Participants who were best fitted by noisy-readout models have quite specific confidence distributions with pronounced probability masses at the extremes and very thin coverage at intermediate confidence levels (participants 4-6, 8, 10, 12, 13, 19) − except those, for which the lognormal readout noise distribution was optimal (participants 11 and 16). Finally, a single participant was best fitted by a censored distribution (participant 14), contrary to the general tendency. This participant likewise had a fairly idiosyncratic confidence distribution characterized by the combination of a probability mass centered at mid-level confidence ratings and a prominent probability mass at a confidence of 1. While a more detailed analysis of individual differences is beyond the scope of this paper, these examples may point to distinct phenotypes of metacognitive noise.

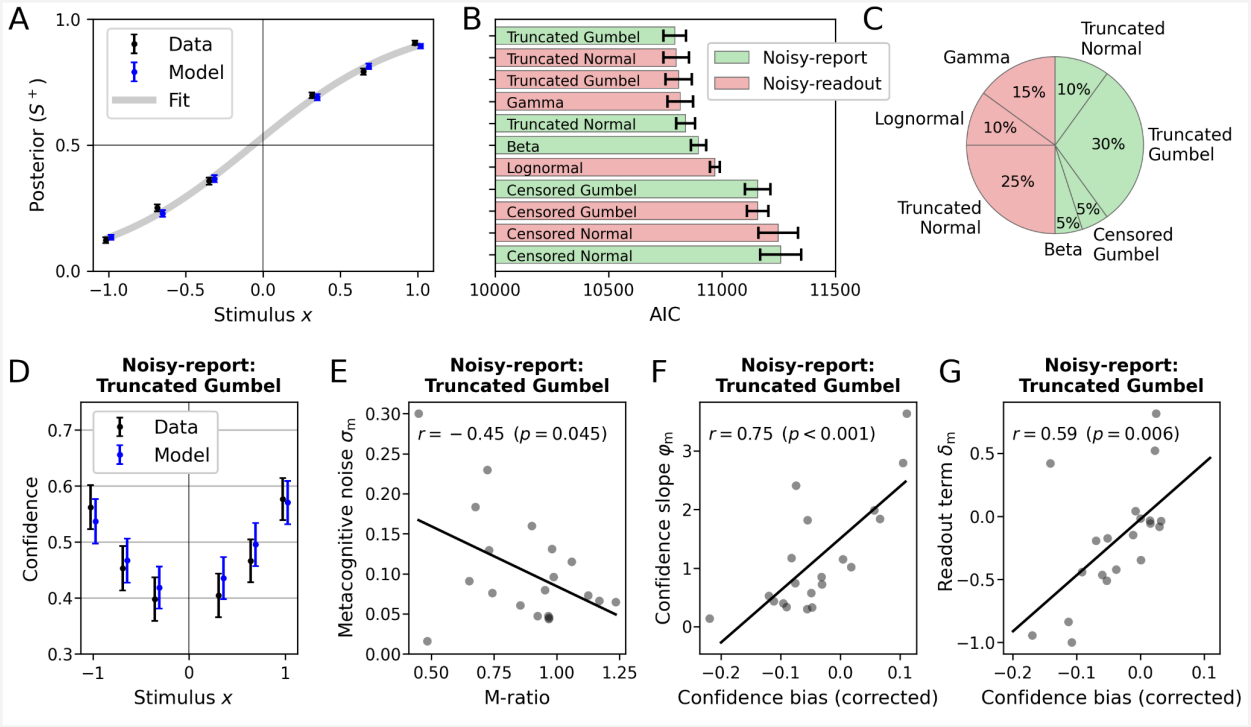

Figure 7. Model performance and relationship between model-based parameters and behavioural measures. (A) Posterior probability (choice probability for S^+) as a function of normalized signed stimulus intensity. Model-based predictions closely follow the empirical data. Means and standard errors (across subjects) were computed for the three difficulty levels of each stimulus category. The fit is based on a logistic function with a bias parameter. **(B)** Akaike information criterion (AIC) for different metacognitive noise distributions. Note that AIC values represent the model evidence of the metacognitive level (sensory levels are identical between models). Error bars indicate standard errors of the mean. **(C)** Breakdown of the best-fitting models for all participants. **(D-G)** Winning model (noisy report + truncated Gumbel). **(D)** Empirical confidence is well-fitted by model-based predictions of confidence. Model-based predictions are based on an average of 1000 runs of the generative model. Error bars represent SEM. **(E)** Relationship of empirical M_{ratio} and model-based metacognitive noise σ_m. **(F)** Relationship of the empirical confidence bias and model-based confidence slope φ_m. The readout term parameter was partialed out from the confidence bias. **(G)** Relationship of the empirical confidence bias and model-based readout term δ_m. The confidence slope parameter was partialed out from the confidence bias.

In the next step, I inspected the winning metacognitive model (noisy report + truncated Gumbel) in more detail. While the selection of this specific model is arbitrary due to the similar performance of several other models, it serves the illustrative purpose and the differences between these models were overall negligible.

I first compared confidence ratings predicted by the model with empirical confidence ratings across the range of experimental stimulus intensities. As shown in Figure 7D, model-predicted confidence tracked behavioral confidence quite well (Figure 7D). This included a slight confidence bias towards S^+, which itself is likely a result of the general sensory bias towards S^+.

I then compared the fitted parameter values of the model with conventional behavioral measures of metacognition. In Result section 1.8, a tight inverse relationship between metacognitive efficiency (M_{ratio})

and the metacognitive noise parameter σ_m was demonstrated for simulated data. As shown in Figure 7E, for the empirical data there was likewise a negative relationship, although weaker ($r_{Pearson} = -0.45$, $p = 0.045$). Note that this relationship is by no means self-evident, as M_{ratio} values are based on information that is not available to a process model: which *specific* responses are correct or incorrect. I will elaborate more on this aspect in the discussion, but assert for now that metacognitive efficiency in empirical data can, at least in part, be accounted for by modeling metacognitive noise in a process model.

As outlined above, the readout term δ_m and confidence slope φ_m can be interpreted as metacognitive biases. To assess the validity of these parameters, I computed individual confidence biases by subtracting a participant's objective accuracy from the subjective predicted accuracy (based on confidence ratings). Positive and negative values of this confidence bias are often regarded as evidence for over- and underconfidence. As shown in Figures 7F and 7G, both parameters show the expected relationships: higher individual confidence biases are associated with higher values of δ_m when controlling for φ_m ($r_{Partial} = 0.75$, $p < 0.001$), and with φ_m when controlling for δ_m ($r_{Partial} = 0.59$, $p = 0.006$). This analysis confirms that the metacognitive bias parameters of the model meaningfully relate to the over- and underconfidence behavior in empirical data.

In a final step, I focus on the model fit of a single participant (Figure 8). The selected participant has a relatively high degree of sensory noise (proportion correct = 0.74, $\sigma_s = 1.04$) compared to the group mean (proportion correct ± SEM = 0.78 ± 0.01, σ_s ± SEM = 0.89 ± 0.04), reflected in a relatively flat psychometric curve (Figure 8A). Due to a negative response bias ($\delta_s = -0.23$) the psychometric curve is shifted upwards. Like many participants in the dataset, the participant thus tends to disproportionally choose clockwise/S^+ over counterclockwise/S^-.

Figures 8B and 8C visualize the results of the metacognitive level, which is again of the type noisy-report + truncated Gumbel. For this participant, the model fit indicates a negative readout term δ_m, thereby introducing a threshold below which decision values *y* are not metacognitively accessible (indicated by a flat region for the link function in Figure 8B). This negative readout term is compensated by a relatively high confidence slope $\varphi_m = 1.15$, resulting in an average confidence of 0.488 that is close to the group average (0.477 ± 0.038).

While below average in terms of type 1 performance, this participant excels in terms of metacognitive performance. This is both indicated by a high M_{ratio} of 1.23 (group mean ± SEM = 0.88 ± 0.05) and a low metacognitive noise parameter σ_m of 0.06 (group mean ± SEM = 0.10 ± 0.02).

It is important to note that a low metacognitive noise parameter σ_m does not imply that the participants' confidence ratings are expected to be within a narrow range for different stimulus intensities. This is because the uncertainty of the sensory level translates to the metacognitive level: the width of decision value distributions, as determined by sensory noise σ_s, also affects the expected width of downstream confidence distributions. Indeed, the behavioral confidence distributions in Figure 8C are spread out

across the entire confidence range for all difficulty levels. In Figure 8C this aspect is emphasized by not only showing the confidence likelihood for the predicted mean $\bar{y}$ of decision values, but also for sensory decision values 0.5 standard deviations below and above $\bar{y}$.

Note that when considering decision values 0.5 standard deviations above $\bar{y}$, a sign flip occurs for the two lower stimulus intensities of S^- (indicated with likelihood distributions shaded in red). In these cases, the participant would make an incorrect choice. The lowest stimulus intensity also shows the paradoxical situation that confidence is predicted to be higher for incorrect choices at $\bar{y} + 0.5SD$ than for correct choices at $\bar{y}$. This is because the predicted confidence for correct choices in the most difficult S^- condition is so low, that relatively small noise fluctuations in decision value space can lead to incorrect choices with higher confidence. In a similar vein, the two lower stimulus intensities of S^- show a well-known characteristic of statistical confidence: an increase of confidence for incorrect choices as stimulus difficulty increases (Sanders et al., 2016).

To compare the empirical confidence distribution of this participant with the distribution predicted by the model, the parameters in the generative model were set to their corresponding fitted values and sampled confidence ratings. The density histograms obtained from this sampling procedure are shown in Figure 8C (orange lines) and demonstrate a close fit with the participant's confidence rating distributions. This close correspondence is not limited to this particular participant. As shown in Supplementary Figure S8, a generative model described by σ_m, δ_m and φ_m is able to approximate a wide range of idiosyncratic empirical confidence distributions.

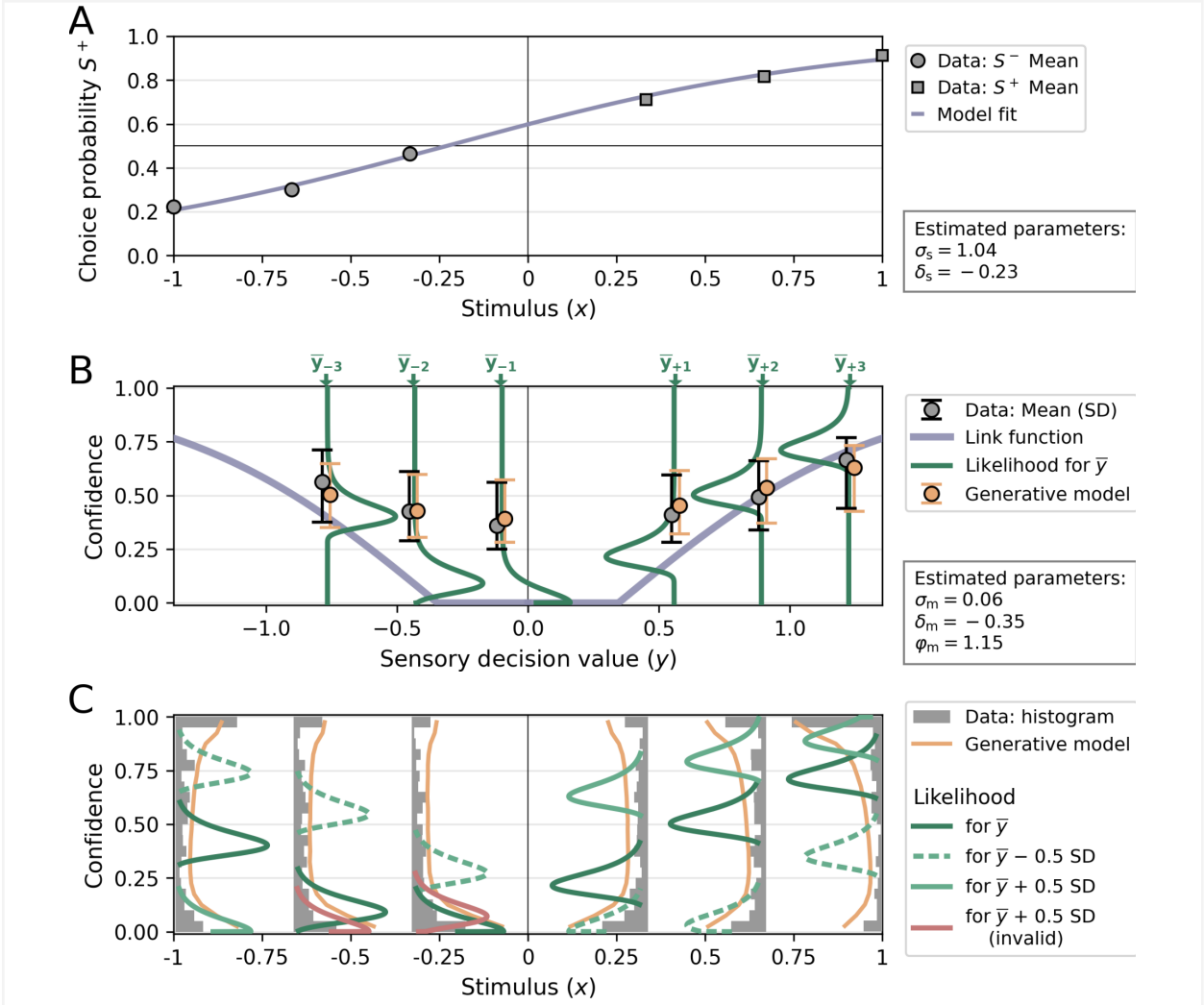

Figure 8. Parameter estimation for a single participant (model: noisy-report + truncated Gumbel). (A) Choice probability for S^+ as a function of stimulus intensity. The negative sensory bias δ_s shifts the logistic function upwards, thereby increasing the choice probability for S^+. **(B)** Link function and confidence likelihood. The link function is specified by the confidence slope φ_m and transforms absolute decision values y (which consider a readout term δ_m) to confidence. The likelihood for confidence ratings is shown only for the mean $\bar{y}_i$ of all possible decision values (separately for the 2 stimulus categories x 3 stimulus intensities). Empirical confidence distributions are indicated by their mean and standard deviation and are plotted at $\bar{y}_i$ for convenience. **(C)** Confidence distributions, likelihood and generative model. Empirical confidence ratings (histograms) and ratings (line plots) obtained from the fitted generative model are anchored at the six levels of the stimulus variable x. To visualize the effect of sensory uncertainty on the metacognitive level, likelihood distributions are plotted not only for the mean $\bar{y}_i$ of the decision value distribution, but also half a standard deviation below (dashed and lighter color) and above (solid and lighter color) the mode. The width of likelihood distributions is controlled by the metacognitive noise parameter σ_m. Distributions colored in red indicate that a sign flip of decision values has occurred, i.e. responses based on these decision values would be incorrect.

Discussion

The present work introduces and evaluates a process model of metacognition and the accompanying toolbox *ReMeta* (see Materials and Methods). The model connects key concepts in metacognition research − metacognitive readout, metacognitive biases, metacognitive noise − with the goal of providing an account of human metacognitive responses. The model can be directly applied to confidence datasets of any perceptual or non-perceptual modality, with the prerequisite that multiple difficulty levels are tested.

As any cognitive computational model, the model can serve several purposes such as inference about model parameters, inference about latent variables and as a means to generate artificial data. In the present work, I focused on parameter inference, in particular metacognitive parameters describing metacognitive noise (σ_m) and metacognitive biases (δ_m, φ_m, λ_m). Indeed, I would argue that this use case is one of the most pressing issues in metacognition research: characterizing the latent processes underlying human confidence reports without the confound of type 1 behavior that hampers descriptive approaches.

In the context of metacognitive biases, I have shown that the conventional method of simply comparing objective and subjective performance (via confidence ratings) is flawed not only because it is biased towards overconfidence, but also because it is strongly dependent on type 1 performance. Just as in the case of metacognitive performance, unbiased inferences about metacognitive biases thus require a process model approach.

Here, I introduced three metacognitive bias parameters which specify metacognitive biases at readout (δ_m), at the transformation from metacognitive evidence to confidence (φ_m) and at report (λ_m). As shown through the simulation of calibration curves, all three parameters can yield under- or overconfidence relative to a bias-free observer. The fact that the calibration curves and the relationships between type 1 performance and confidence biases are quite distinct between the three proposed metacognitive bias parameters may indicate that these are to some degree dissociable. Moreover, in an empirical dataset the readout term δ_m and the confidence slope φ_m strongly correlated with a conventional confidence bias measure, further validating these parameters.

The second kind of metacognitive parameter considered in this work is metacognitive noise (Mueller and Weidemann, 2008; Jang et al., 2012; De Martino et al., 2013; van den Berg et al., 2017; Bang et al., 2019; Shekhar and Rahnev, 2021). As with metacognitive biases, metacognitive noise may arise at different stages of the processing hierarchy and in the present work I investigated two kinds: noise at readout and report. Both parameters affect the precision of confidence ratings and as a result they showed an expected negative relationship with regular measures of metacognitive ability (*AUROC2*, M_{ratio}). Importantly, I show that while even M_{ratio}, a measure normalized for type 1 performance, was dependent on type 1 performance for simulated data, recovered estimates of metacognitive noise were largely

invariant to type 1 performance. Thus, just as in the case of metacognitive biases, the entanglement of metacognitive and type 1 behavior can be unraveled by means of a process model approach.

While this summary so far emphasized the advantages of a process model approach to metacognition, there are a number of remaining challenges. First, it is entirely possible that a comprehensive model of metacognition is non-invertible from the point of confidence ratings. This challenge is exemplified by the noisy-readout model, for which the inversion requires a strictly monotonically increasing link function. The cure, in such a scenario, would be to obtain additional measures along the processing hierarchy. For instance, reaction time could be considered an implicit proxy for confidence, which is affected by readout noise but not by reporting noise. Practically, this amounts to inserting a separate level of implicit metacognition between the sensory and the (explicit) metacognitive level. Implicit metacognitive noise and bias parameters would be obtained by maximizing the likelihood of measured reaction times and explicit metacognitive parameters by maximizing the likelihood of confidence ratings with respect to the output of the implicit level.

Second, the effects of different sources of bias and noise along the processing hierarchy may be so strongly correlated that their dissociation would require unreachable amounts of confidence data. This dissociation, however, is essential for many research questions in metacognition – whether the goal is to derive a fundamental model of human metacognition or whether one is interested in specific abberrancies in mental illness. An example for the latter is the frequent observation of overconfidence in schizophrenia which is thought to reflect a more general deficit in the ability to integrate disconfirmatory evidence (Speechley, 2010; Zawadzki et al., 2012) and may underlie the maintenance of delusional beliefs (Moritz and Woodward, 2006b). To investigate this specific hypothesis, it is central to dissociate whether metacognitive biases mainly apply at the reporting stage – which may be a result of the disease – or at an earlier metacognitive processing stage, which may be causally involved in the development of the disease. This issue likewise could be addressed by measuring behavioral, physiological or neurobiological processes that precede the report of confidence.

Third, the demonstration of an unbiased recovery of metacognitive noise and bias parameters in a process model approach comes with a strong caveat, since the data is generated with the very same model that is used for parameter recovery. Yet, all models are wrong, goes the saying, and this certainly applies to current models of metacognition. The question is thus: given the unknown true model/s that underlie/s empirical confidence ratings, to what degree can parameters obtained from an approximated model be considered unbiased? The way forward here is to continuously improve computational models of metacognition in terms of model evidence, thus increasing the chances that fitted parameters are meaningful estimates of the true parameters.

With respect to previous modeling work, a recent paper by Shekhar and Rahnev (2021) deserves special attention. Here too, the authors adopted a process model approach for metacognition with the specific goal of deriving a measure of metacognitive ability, quite similar to the metacognitive noise parameter

σ_m in this work. One key difference is that Shekar and Rahnev tailored their model to discrete confidence scales, such that each possible confidence rating (for each choice option) is associated with a separately fitted confidence criterion. This introduces maximal flexibility, as essentially arbitrary mappings from internal evidence to confidence can be fitted. In addition, it requires minimal assumptions about the link functions that underlies the computation of confidence, apart from an ordering constraint applied to the criteria.

However, while this flexibility is a strength, it also comes at certain costs. One issue is the relatively large number of parameters that have to be fitted. Shekar and Rahnev note that the MLE procedures for the fitting of confidence criteria often got stuck in local minima. Rather than via MLE, confidence criteria were thus fitted by matching the expected proportion of high confidence trials to the observed proportion for each criterion. It is thus not guaranteed that the obtained confidence criterions indeed maximize the likelihood under the data. Furthermore, to make a criterion-based model compatible with data from a continuous confidence scale, confidence reports have to be discretized. Apart from the loss of information associated with discretization, this introduces uncertainty as to how exactly the data should be binned (e.g. equinumerous versus equidistant). Another aspect worth mentioning is that a criterion-based approach effectively corresponds to a stepwise link function, which is not invertible. Making inferences about readout noise thus poses a challenge to such criterion-based models.

In the present work, I assumed a mapping between internal evidence and confidence that can be described by a parametric link function. This too comes with advantages and disadvantages. On the one hand, a parametric link function naturally imposes strong constraints on the mapping between internal evidence and confidence. In reality, this mapping might not conform to any simple function – and even if it did, different observers might apply different functions. On the other hand, imposing a specific link function can be seen as a form of regularization when statistical power is insufficient to constrain a large number of individual criteria. Further, a parametric link does not need to worry about the discretization of confidence ratings, while still being compatible with a priori discretized ratings. And finally, a meaningful inference about metacognitive biases requires a parametric link function which computes the subjective probability of being correct (as in Equation 5).

The disproportionate frequency of the most extreme ratings for continuous confidence scales – both in the empirical dataset considered here (Shekhar and Rahnev, 2021), but also in many other datasets (Rahnev et al., 2020) – may indicate that there are at least two criteria at play: a finite level of metacognitive evidence below which the lowest confidence rating is given (minimum confidence criterion) and a level of metacognitive evidence above which the highest confidence rating is given (maximum confidence criterion). Yet, even without the postulation of explicit criteria, the present model can account in two possible, non-mutually exclusive, ways for this observation. The first possibility is that extremal ratings are a consequence of metacognitive noise, which pushes confidence ratings to the

bounds. In particular, censored noise distributions can readily account for an accumulation of confidence ratings at the bounds of the scale.

The second explanation is that the minimum and maximum confidence criteria implicitly arise as a consequence of metacognitive biases (Result section 1.4). Indeed, the individual fits to the empirical dataset (Supplementary Figure S8) showed that the model accounted well even for data of participants with a strong preference for extremal confidence ratings, despite not considering explicit confidence criteria. An exploratory analysis (not reported here) indicated that participants with such confidence distributions are not well-explained by censored noise models, which may be a tentative argument for the metacognitive bias explanation instead.

The process model approach deviates in an important way from standard analyses of confidence reports based on the type 2 receiver operating curve. As type 2 ROC analyses are solely based on stimulus-specific type 1 and type 2 responses, they do not consider one of the arguably most important factors in this context: stimulus intensity. This implies that such measures cannot dissociate to what degree variability in confidence ratings is based on stimulus variability or on internal noise. In contrast, since a process model specifies the exact transformation from stimulus intensity to decision variable to confidence, this source of variance is appropriately taken into account. The metacognitive noise parameter σ_m introduced here is thus a measure of the *unexpected* variability of confidence ratings, after accounting for the variability on the stimulus side.

Yet, the process model approach bears another important difference compared with type 2 ROC analyses, in this case a limiting factor on the side of the process model. As the area under the type 2 ROC quantifies to what degree confidence ratings discriminate between correct and incorrect responses, it is important to recognize what valuable piece of information the correctness of a *specific* response is. Over and above stimulus intensity, the correctness of a response will typically be influenced by negative factors such as attentional lapses, finger errors, tiredness, and positive factors such as phases of increased motivation or concentration. All of these factors not only influence type 1 performance, but they also influence the type 2 response that one would expect from an ideal metacognitive observer. Analyses of type 2 ROCs implicitly make use of this information insofar as they consider the correctness of each individual response.

In contrast, this information is not available in a process model. The signal that enters the metacognitive level of the process model is based on a prediction that is computed solely on the basis of experimental factors such as stimulus difficulty, but not based on the correctness of specific choices. Note that this is not a limitation specific to the present model; it is the nature of a process model that it makes predictions based on inputs to the system and not based on (parts of) the system outputs it aims to predict. Improving process models in this regard necessitates access to additional trial-by-trial data based on objective measurements (e.g. eye-tracking data) or subjective reports if these are not used for model fitting (e.g. reports of attentional lapses or finger errors).

Conclusion

The model outlined in this paper casts confidence as a noisy and potentially biased transformation of sensory decision values. The model parameters that shape this transformation provide a rich account of human metacognitive inefficiencies and metacognitive biases. In particular, I hope that the underlying framework will allow a systematic model comparison in future confidence datasets to elucidate sources of metacognitive noise, to narrow down candidate noise distributions and to differentiate between different kinds of metacognitive biases. The accompanying toolbox *ReMeta* provides the functionality for such investigations.

Materials and Methods

The ReMeta toolbox

The code underlying this work has been bundled in a user-friendly Python toolbox (*ReMeta*) which is published alongside this paper at github.com/m-guggenmos/remeta. While its core is identical to the framework outlined here, it offers a variety of additional parameters and settings. In particular, it allows fitting separate values for each parameter depending on the sign of the stimulus (for sensory parameters) or the decision value (for metacognitive parameters). Moreover, it offers various choices for noise distributions and link functions, including criterion-based link functions.

The *ReMeta* toolbox has a simplified interface such that in the most basic case it requires only three data vectors as input: stimuli, choices and confidence. The output is a structure containing the fitted parameters, information about the goodness of fit (log-likelihood, AIC, BIC, correlation between empirical confidence ratings and ratings from a generative model) and optionally vectors of all latent variables (e.g. decision values, metacognitive evidence). The toolbox is highly configurable – in particular each parameter can be disabled, enabled or enabled in duplex mode (i.e. sign-dependent, see above).

Parameter fitting minimizes the negative log-likelihood of type 1 choices (sensory level) or type 2 confidence ratings (metacognitive level). For the sensory level, initial guesses for the fitting procedure were found to be of minor importance and are thus set to reasonable default values (0.1 for sensory noise σ_s and 0 for sensory bias δ_s and threshold ϑ_s). Data are generally fitted with a gradient-based optimization method (*Sequential Least Squares Programming*; Kraft, 1988). However, if enabled, the sensory threshold parameter can introduce a discontinuity in the psychometric function, thereby violating the assumptions of gradient methods. In this case, an additional gradient-free method (*Powell's method*; Powell, 1964) is fitted and the estimate with the lower negative log-likelihood is chosen. Both parameter fitting procedures respect lower and upper bounds for each parameter.

Since parameters of the metacognitive level were found to be more variable, subject-specific initial values for the fitting procedure are of greater importance. For this reason, an initial coarse grid-search with parameter-specific grid points is performed to determine suitable initial values, which are

subsequently used for a gradient-based optimization routine (*Sequential Least Squares Programming*). The combined procedure likewise respects lower and upper bounds for each parameter.

All analyses and simulations in this work were performed with these default settings. In addition, the toolbox has optional settings to invoke an additional fine-grained grid-search and an explicit global optimization routine (*Basin-hopping*; Wales and Doye, 1997), both of which are computationally considerably more expensive. Exploratory tests showed that these methods were not necessary for parameter estimation on either simulated or empirical data in this work; however, this may be different for other empirical datasets.

Acknowledgements

This research was funded by the German Research Foundation (grant GU 1845/1-1). I'm grateful to the lab of Elisa Filevich for helpful input and critical discussion. Computation has been performed on the HPC for Research cluster of the Berlin Institute of Health.

Competing interests

The author declares that no competing interests exist.

References

Adler WT, Ma WJ (2018) Limitations of Proposed Signatures of Bayesian Confidence. Neural Comput 30:3327–3354.

Akaike H (1974) Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6), 716–723. doi:10.1109/tac.1974.1100705. IEEE Trans Autom Control 19:716–723.

Bang JW, Shekhar M, Rahnev D (2019) Sensory noise increases metacognitive efficiency. J Exp Psychol Gen 148:437–452.

Clarke FR, Birdsall TG, Tanner WP (1959) Two Types of ROC Curves and Definitions of Parameters. J Acoust Soc Am 31:629–630.

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

Dosher BA, Lu ZL (1998) Perceptual learning reflects external noise filtering and internal noise reduction through channel reweighting. Proc Natl Acad Sci U S A 95:13988–13993.

Drugowitsch J (2016) Becoming Confident in the Statistical Nature of Human Confidence Judgments. Neuron 90:425–427.

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

Fleming SM, Daw ND (2017) Self-Evaluation of Decision-Making: A General Bayesian Framework for Metacognitive Computation. Psychol Rev 124:91–114.

Fleming SM, Lau H (2014) How to measure metacognition. Front Hum Neurosci 8:443.

Fleur DS, Bredeweg B, van den Bos W (2021) Metacognition: ideas and insights from neuro- and educational sciences. Npj Sci Learn 6:13.

Fu T, Koutstaal W, Fu CHY, Poon L, Cleare AJ (2005) Depression, confidence, and decision: Evidence against depressive realism. J Psychopathol Behav Assess 27:243–252.

Fu T, Koutstaal W, Poon L, Cleare AJ (2012) Confidence judgment in depression and dysphoria: The depressive realism vs. negativity hypotheses. J Behav Ther Exp Psychiatry 43:699–704.

Fullerton G, Cattell J (1892) On the Perception of Small Differences: With Special Reference to the Extent, Force and Time of Movement. Philadelphia, PA: University of Pennsylvania Press.

Galvin SJ, Podd JV, Drga V, Whitmore J (2003) Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. Psychon Bull Rev 10:843–876.

Gigerenzer G, Hoffrage U, Kleinbfilting H (1991) Probabilistic Mental Models: A Brunswikian Theory of Confidence. Psychol Rev 98:506–528.

Guggenmos M (2021) Validity and reliability of metacognitive performance measures. BioRxiv.

Hangya B, Sanders JI, Kepecs A (2016) A Mathematical Framework for Statistical Decision Confidence. Neural Comput 28:1840–1858.

Harvey N (1997) Confidence in judgment. Trends Cogn Sci 1:78–82.

Hoven M, Lebreton M, Engelmann JB, Denys D, Luigjes J, van Holst RJ (2019) Abnormalities of confidence in psychiatry: an overview and future perspectives. Transl Psychiatry 9:268.

Jang Y, Wallsten TS, Huber DE (2012) A Stochastic Detection and Retrieval Model for the Study of Metacognition. Psychol Rev 119:186–200.

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

Köther U, Veckenstedt R, Vitzthum F, Roesch-Ely D, Pfueller U, Scheu F, Moritz S (2012) “Don’t give me that look” - Overconfidence in false mental state perception in schizophrenia. Psychiatry Res 196:1–8.

Kraft D (1988) A software package for sequential quadratic programming. German Aerospace Center.

Lichtenstein S, Fischhoff B (1977) Do those who know more also know more about how much they know? Organ Behav Hum Perform 20:159–183.

Lichtenstein S, Fischhoff B, Phillips L (1977) Calibration of probabilities: The state of the art. Decis Mak Change Hum Aff:275–324.

Lichtenstein S, Fischhoff B, Phillips LD (1982) Calibration of probabilities: The state of the art to 1980. In: Judgment under uncertainty (Kahnemann D, Slovic P, Tversky A, eds), pp 306–334. Cambridge, UK: Cambridge University Press.

Maniscalco B, Lau H (2012) A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. Conscious Cogn 21:422–430.

Maniscalco B, Lau H (2014) Signal Detection Theory Analysis of Type 1 and Type 2 Data: Meta-d0, Response- Specific Meta-d0, and the Unequal Variance SDT Model. In: The cognitive neuroscience of metacognition (Fleming SM, Frith CD, eds), pp 25–66. Basel: Springer-Verlag Publishing. Available at: https://doi.org/10.1007/978-3-642-45190-4_3.

Maniscalco B, Lau H (2016) The signal processing architecture underlying subjective reports of sensory awareness. Neurosci Conscious 2016 Available at: https://academic.oup.com/nc/article/doi/10.1093/nc/niw002/2757122 [Accessed August 1, 2021].

Merkle EC (2009) The disutility of the hard-easy effect in choice confidence. Psychon Bull Rev 16:204–213.

Moritz S, Lysaker PH (2019) Metacognition Research in Psychosis: Uncovering and Adjusting the Prisms That Distort Subjective Reality. Schizophr Bull 45:17–18.

Moritz S, Ramdani N, Klass H, Andreou C, Jungclaussen D, Eifler S, Englisch S, Schirmbeck F, Zink M (2014) Overconfidence in incorrect perceptual judgments in patients with schizophrenia. Schizophr Res Cogn 1:165–170.

Moritz S, Woodward TS (2006a) The contribution of metamemory deficits to schizophrenia. J Abnorm Psychol 115:15–25.

Moritz S, Woodward TS (2006b) Metacognitive control over false memories: A key determinant of delusional thinking. Curr Psychiatry Rep 8:184–190.

Mueller ST, Weidemann CT (2008) Decision noise: An explanation for observed violations of signal detection theory. Psychon Bull Rev 15:465–494.

Nelson TO (1984) A Comparison of Current Measures of the Accuracy of Feeling-of-Knowing Predictions. Psychol Bull 95:109–133.

Pierce CS, Jastrow J (1885) On small differences of sensation. Mem Natl Acad Sci 3:73–83.

Pollack I (1959) On Indices of Signal and Response Discriminability. J Acoust Soc Am 31:1031–1031.

Powell MJD (1964) An efficient method for finding the minimum of a function of several variables without calculating derivatives. Comput J 7:155–162.

Rahnev D et al. (2020) The Confidence Database. Nat Hum Behav.

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

Rounis E, Maniscalco B, Rothwell JC, Passingham RE, Lau H (2010) Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. Cogn Neurosci 1:165–175.

Rouy M, Saliou P, Nalborczyk L, Pereira M, Roux P, Faivre N (2020) Systematic review and meta-analysis of the calibration of confidence judgments in individuals with schizophrenia spectrum disorders. medRxiv Available at: http://medrxiv.org/lookup/doi/10.1101/2020.12.03.20243113 [Accessed January 13, 2021].

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

Seow TXF, Rouault M, Gillan CM, Fleming SM (2021) How Local and Global Metacognition Shape Mental Health. Biol Psychiatry Available at: https://www.sciencedirect.com/science/article/pii/S0006322321013299 [Accessed August 4, 2021].

Shekhar M, Rahnev D (2021) The nature of metacognitive inefficiency in perceptual decision making. Psychol Rev 128:45–70.

Soll JB (1996) Determinants of Overconfidence and Miscalibration: The Roles of Random Error and Ecological Structure. Organ Behav Hum Decis Process 65:117–137.

Speechley W (2010) The contribution of hypersalience to the “jumping to conclusions” bias associated with delusions in schizophrenia. J Psychiatry Neurosci 35:7–17.

van den Berg R, Yoo AH, Ma WJ (2017) Fechner’s law in metacognition: A quantitative model of visual working memory confidence. Psychol Rev 124:197–214.

Wales DJ, Doye JPK (1997) Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms. J Phys Chem A 101:5111–5116.

Xue K, Shekhar M, Rahnev D (2021) The nature of metacognitive noise confounds metacognitive sensitivity and metacognitive bias. psyArXiv Available at: https://osf.io/buahk [Accessed March 4, 2021].

Zawadzki JA, Woodward TS, Sokolowski HM, Boon HS, Wong AHC, Menon M (2012) Cognitive factors associated with subclinical delusional ideation in the general population. Psychiatry Res 197:345–349.