RESEARCH ARTICLE

Explaining distortions in metacognition with an attractor network model of decision uncertainty

Nadim A. A. Atiya[1,2]*, **Quentin J. M. Huys**[1,3], **Raymond J. Dolan**[1,2], **Stephen M. Fleming**[1,2,4]

1 Max Planck University College London Centre for Computational Psychiatry and Ageing Research, University College London, London, United Kingdom, **2** Wellcome Centre for Human Neuroimaging, University College London, London, United Kingdom, **3** Division of Psychiatry, University College London, London, United Kingdom, **4** Department of Experimental Psychology, University College London, London, United Kingdom

* n.atiya@ucl.ac.uk

Abstract

Metacognition is the ability to reflect on, and evaluate, our cognition and behaviour. Distortions in metacognition are common in mental health disorders, though the neural underpinnings of such dysfunction are unknown. One reason for this is that models of key components of metacognition, such as decision confidence, are generally specified at an algorithmic or process level. While such models can be used to relate brain function to psychopathology, they are difficult to map to a neurobiological mechanism. Here, we develop a biologically-plausible model of decision uncertainty in an attempt to bridge this gap. We first relate the model's uncertainty in perceptual decisions to standard metrics of metacognition, namely mean confidence level (bias) and the accuracy of metacognitive judgments (sensitivity). We show that dissociable shifts in metacognition are associated with isolated disturbances at higher-order levels of a circuit associated with self-monitoring, akin to neuropsychological findings that highlight the detrimental effect of prefrontal brain lesions on metacognitive performance. Notably, we are able to account for empirical confidence judgements by fitting the parameters of our biophysical model to first-order performance data, specifically choice and response times. Lastly, in a reanalysis of existing data we show that self-reported mental health symptoms relate to disturbances in an uncertainty-monitoring component of the network. By bridging a gap between a biologically-plausible model of confidence formation and observed disturbances of metacognition in mental health disorders we provide a first step towards mapping theoretical constructs of metacognition onto dynamical models of decision uncertainty. In doing so, we provide a computational framework for modelling metacognitive performance in settings where access to explicit confidence reports is not possible.

Citation: Atiya NAA, Huys QJM, Dolan RJ, Fleming SM (2021) Explaining distortions in metacognition with an attractor network model of decision uncertainty. PLoS Comput Biol 17(7): e1009201. https://doi.org/10.1371/journal.pcbi.1009201

Editor: Alireza Soltani, Dartmouth College, UNITED STATES

Received: September 25, 2020

Accepted: June 18, 2021

Published: July 26, 2021

Data Availability Statement: Code used to fit, simulate, and analyse the model (and data) is available at this repository: https://github.com/nidstigator/uncertainty_psychiatry_2020 Data collected is available in the same repository.

Funding: SMF is funded by a Wellcome/Royal Society Sir Henry Dale Fellowship (206648/Z/17/Z) and a Philip Leverhulme Prize from the Leverhulme Trust. The Wellcome Centre for Human Neuroimaging is supported by core funding from the Wellcome Trust (206648/Z/17/Z). The Max Planck UCL Centre is a joint initiative supported by

UCL and the Max Planck Society. The funders did not play any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

In this work, we use a biologically-plausible model of decision uncertainty to show that shifts in metacognition are associated with disturbances in the interaction between decision-making and higher-order uncertainty-monitoring networks. Specifically, we show that uncertainty modulation is associated with metacognitive bias, sensitivity, and efficiency, with no effect on perceptual sensitivity. Our approach not only enables inferences about uncertainty modulation (and, in turn, these facets of metacognition) from fits to first-order performance data alone–but also provides a first step towards relating dynamical models of decision-making to metacognition. We also relate our model's uncertainty modulation to psychopathology, and show that it can offer an implicit, low-dimensional marker of metacognitive (dys)function–opening the door to richer analysis of the interaction between metacognitive performance and psychopathology from first-order performance data.

Introduction

Computational psychiatry [1–4] employs mechanistic and theory-driven models to relate brain function to phenomena that characterise mental health disorders [2,5–8]. Typically, algorithmic-level models [9] describe the computational processes that realise specific brain functions and return theoretically meaningful parameters that may vary between subjects. Some of these algorithmic models (e.g. reinforcement learning; [8]) closely relate to the functions of discrete brain circuits [10–12]. However, there remains a high degree of imprecision when relating diverse sets of algorithms to circuit-level disturbances, potentially limiting our understanding of, and treatments for, mental disorders.

One proposal is that the same neural circuit disturbances can be associated with several (often unrelated) changes in behaviour [13]. Here detailed biophysical models [14–16] may provide tools for understanding mental health disorders in terms of precise disturbances at the microcircuit level. For instance, [14] showed that an imbalance in excitatory/inhibitory synaptic connections in a spiking neural network model can explain working memory deficits associated with schizophrenia. However, the complex nature of such models renders it challenging to fit them to individual subjects' behavioural data. At the level of neural systems, simpler biologically-grounded models [17,18] have been employed to relate macrocircuit-level dysfunctions to symptoms of mental health disorders, and motivate non-invasive experimental neuroimaging to probe such dysfunctions [19]. Such (connectionist) biologically-motivated models retain a mapping between neurobiology and behaviour, while allowing faster computation and fewer free parameters.

Here our focus is on developing similar biologically-plausible models of subjective confidence and metacognition–the ability to reflect upon and evaluate aspects of our own cognition and behaviour. Recent advances in metacognition research has led to the development of precision assays for different facets of metacognitive ability [20,21]. Within a signal detection theory (SDT) framework, metacognitive bias refers to a subject's overall (mean) confidence level on a task. In contrast, metacognitive sensitivity refers to whether subjects' confidence ratings effectively distinguish between correct and incorrect decisions, as quantified by the SDT metric $meta{-}d'$. Furthermore, metacognitive sensitivity can be compared to another SDT measure, d', which quantifies how effectively a subject processes information related to the task [22,23]. The ratio $meta{-}d'/d$ thus yields a measure of metacognitive efficiency, i.e. metacognitive sensitivity for a given level of task performance [24].

Experimental evidence suggests that these facets of metacognitive ability are dissociable from task performance, and may have a distinct neural and computational basis [25–31]. Interestingly, self-reported mental health symptoms have been linked to changes in metacognition, often in the absence of differences in task performance [32–35]. Developing a biologically-motivated model of metacognition has the potential to cast light on how this dissociable mechanism is implemented at a circuit level, as well as provide a direct bridge between circuit-level dysfunction and psychopathology.

Theoretical work addressing perceptual decision-making has proposed dynamical reduced accounts [36,37] that provide detailed biophysical models of decision making [38], enabling more rigorous theoretical analyses and faster computation. For instance, [36] have accounted for most of the behavioural results addressed by [38] using the two slowest N-Methyl-D-aspartic acid (NMDA) dynamical variables. More recently, [39] extended [36] to account for decision confidence reports and other metacognitive behaviours, such as an ability to flexibly change one's mind and correct errors [40]. More specifically, guided by neurophysiological evidence that supports an encoding of confidence within higher-order prefrontal brain regions [27,41], the authors introduced the idea of a third 'uncertainty-monitoring' neuronal population (i.e. dynamical variable). This population continuously monitors uncertainty in the network, interacting with the other two populations involved in decision-making via a feedback loop mechanism [42].

A classic proposal from cognitive psychology is that changes in metacognition reflect alterations in higher-order computations that serve to "monitor" first-order task performance [43]. Our primary focus here is on the question of whether developing biologically-plausible accounts of metacognitive monitoring can shed light on the source of differences metacognitive sensitivity. Other recent work has focused on simulating parallel neural populations engaged in perceptual decision-making, finding that informing confidence with the activity of less-normalisation-tuned neurons can account for cases in which confidence is altered in the absence of differences in performance [44]. Our model is complementary to this endeavour, instead focusing on the dynamics of uncertainty encoding within a dedicated, higher-order neural population that integrates input from sensorimotor neuronal pools, and continuously feeds this uncertainty signal back to modulate evidence integration. This feedback mechanism adds a layer of nonlinearity, accounting for non-trivial interactions between confidence, accuracy and response times. We will see, though, that such a higher-order monitoring population can also account for shifts in metacognitive bias, and therefore capture instances of performance-confidence dissociation.

To gain insight into potential mechanisms underlying shifts in metacognition, we first demonstrate that our biologically-motivated model [39,40] can account for human confidence reports. In a novel approach, we show that the intrinsic dynamics of this model, constrained only by first-order performance, are sufficient to account for subjects' confidence reports, going beyond existing methods of fitting models directly to empirical confidence data [45,46]. We then map theoretical constructs such as metacognitive sensitivity and efficiency onto our dynamical model, demonstrating that changes in metacognitive sensitivity are associated with isolated disturbances in a higher-order node of the network involved in uncertainty monitoring. This computational approach also allowed us to relate circuit-level deficits in metacognition to psychopathology, by re-analysing an existing dataset [32]. We hope our work advances the field by providing a computational framework for mapping theoretical metrics of metacognition onto dynamical models of decision uncertainty.

Results

Neural circuit model

Our model comprises two interacting subnetworks. The sensorimotor module comprises two mutually-inhibiting neuronal populations selective for two decision alternatives (eg more dots on the right or left), each of which are endowed with self-excitation [36]. Importantly, our model builds on neurophysiological evidence suggesting that decision confidence is encoded by higher-order brain regions [27,41]. A crucial aspect of the model is that decision uncertainty (i.e. reciprocal of confidence) is continuously monitored by a dedicated neuronal population termed the 'uncertainty-monitoring' population. The latter encodes uncertainty using a leaky integrator–by integrating the summed neuronal activities of sensorimotor populations (see Fig 1C for a sample trial). This integration is terminated when a response is made, i.e. in effect corresponding to when neuronal activity in one of the sensorimotor populations reaches a decision threshold (see Fig 1C and Methods). Finally, the uncertainty-monitoring population continuously feeds back the encoded uncertainty into both sensorimotor populations via a feedback loop (See Fig 1B, red arrows). This excitatory feedback mechanism is reminiscent of a dynamic gain modulation (see Fig 1D), previously shown to account well for response time patterns from decision-making experiments with urgency [47–51]. Here we refer to this feedback loop as the strength of uncertainty modulation (UM).

Applying the model to account for facets of metacognition

We first asked whether our model can account for variation in standard theoretical metrics of metacognition. To do that, we simulated the model using various parameter values, and derived both choices and confidence judgements from the fluctuations in the uncertainty-monitoring population of the model. More specifically, for each simulated trial, we define decision uncertainty (the inverse of decision confidence) as the maximum firing rate reached by the uncertainty-monitoring population within that trial [39]. We use equal-width binning to bin (discretise) raw confidence measurements into confidence bins (discrete ratings).

Next, we entered the simulated confidence-accuracy matrix as data into a Bayesian model of metacognitive sensitivity [21]. The model returns a parameter *meta–d′* representing the metacognitive sensitivity for a particular simulation with a set of parameter values. Metacognitive efficiency is then estimated by comparing *meta–d′* to the model's perceptual sensitivity (i.e, *d′*) yielding the ratio meta_d'/d' (M-Ratio [20]). Metacognitive bias is defined as the average binned confidence level across both correct and incorrect trials. We fitted several linear models to estimate the contribution of each parameter in our network model of decision confidence to perceptual sensitivity, metacognitive bias, metacognitive sensitivity, and metacognitive efficiency (see Methods).

The results (Fig 2A and 2C) show increasing gain has a strong positive effect on *d′* and metacognitive sensitivity. The effect on *d′* is unsurprising given that increasing gain magnifies the difference in input each neuronal population is receiving (see Fig 1B). *d′* here acts as a ceiling for metacognitive sensitivity, hence the increase in *meta_d'* with increasing gain. Notably, however, we also ran simulations with higher UM values, and metacognitive sensitivity worsened and did not increase with increasing gain, despite the improvement in perceptual sensitivity (Fig D in S5 Appendix). The results also show (Fig 2B) that increasing gain has a weak effect on metacognitive bias (although see Fig D in S5 Appendix). Finally, the results (Fig 2D) show that increasing gain has a moderate positive effect on metacognitive efficiency, possibly driven by the sustained linear increase in *d'* as a function of gain.

More interestingly, the second set (Fig 2, bottom row) of results show that increasing UM has only weak effects on first-order task performance (*d′*) (Fig 2E). However, increasing UM

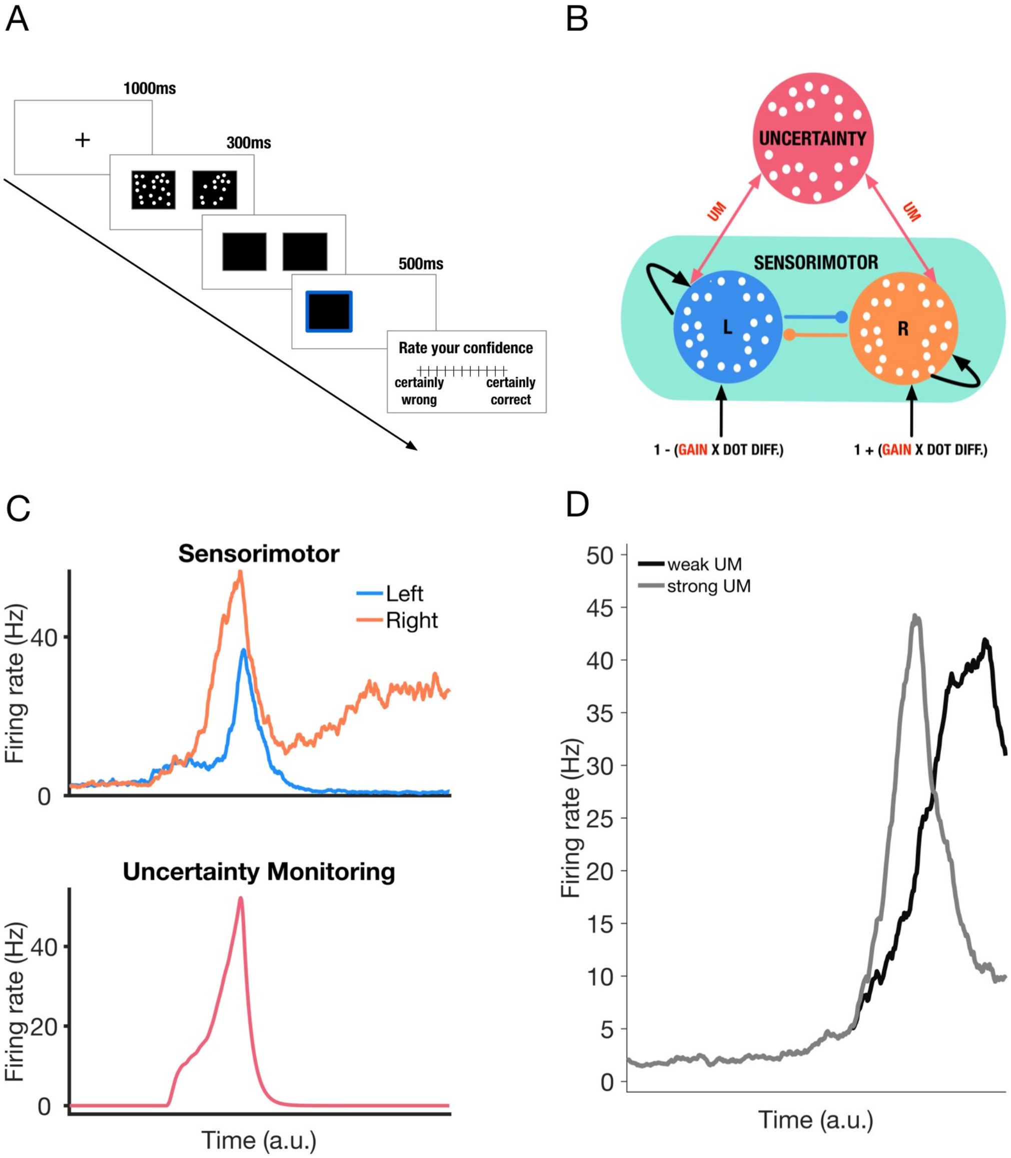

Fig 1. Task and neural circuit model. A. Perceptual decision-making task used as a basis for simulations. A fixation cross appears for 1000ms, followed by two boxes with dots for a fixed duration of 300ms. Subjects are asked to judge which box contains the greater number of dots by pressing left/right key on the keyboard. Their response is highlighted for 500ms, i.e. with a blue border appearing around the chosen box. Finally, participants report their confidence in their decision on a scale of 1–11 in experiment 1, and 1–6 in experiment 2 (S1 and S2 Appendices). **B.** Neural circuit model of decision uncertainty. The model comprises two modules. The sensorimotor module (green) comprises two neuronal populations (blue/orange) selective for right/left information. The two populations are endowed with mutual inhibition (lines with filled circles) and self-excitation (curved arrows). These populations receive external input as a function of the difference between the number of dots shown in the two boxes. Figure assumes correct response is on the right–hence the positive input bias for the population selective to rightward information. A gain parameter controls the difference in input each population receives. One neuronal population (red) continuously monitors overall decision uncertainty by integrating the summed output of the sensorimotor populations (see Methods). Uncertainty is equally fed back into both neuronal populations through symmetric feedback excitation (two-way red arrows, controlled by value of uncertainty modulation strength, UM). **C.** A sample timecourse of the activities of the sensorimotor populations (top panel) and uncertainty-monitoring population (bottom panel). Typical winner-take-all behaviour is seen in the sensorimotor module. Activity of the uncertainty-monitoring population follows a phasic profile (see [39,40] and Methods). Trial simulated with dot difference between the two boxes set at 20 (see Methods). **D.** Sample timecourse of firing rates of the 'winning' neural population (i.e. one with more input bias) in the sensorimotor module under two strengths of uncertainty-modulation (UM) values. Random seed reset to control for noise. In the case of the trial with strong (weak) excitatory feedback (solid grey

(black) trace), ramping up is faster (slower), leading to a quicker (slower) response. Neural population firing rates shown here are smoothed with a simple moving average (window size = 50ms).

https://doi.org/10.1371/journal.pcbi.1009201.g001

strength has a negative effect on both metacognitive bias (Fig 2F) and *meta*–*d*′ (Fig 2G), leading to reductions in overall confidence and metacognitive sensitivity. Given that first-order performance is relatively unchanged, greater UM strength also results in lower metacognitive

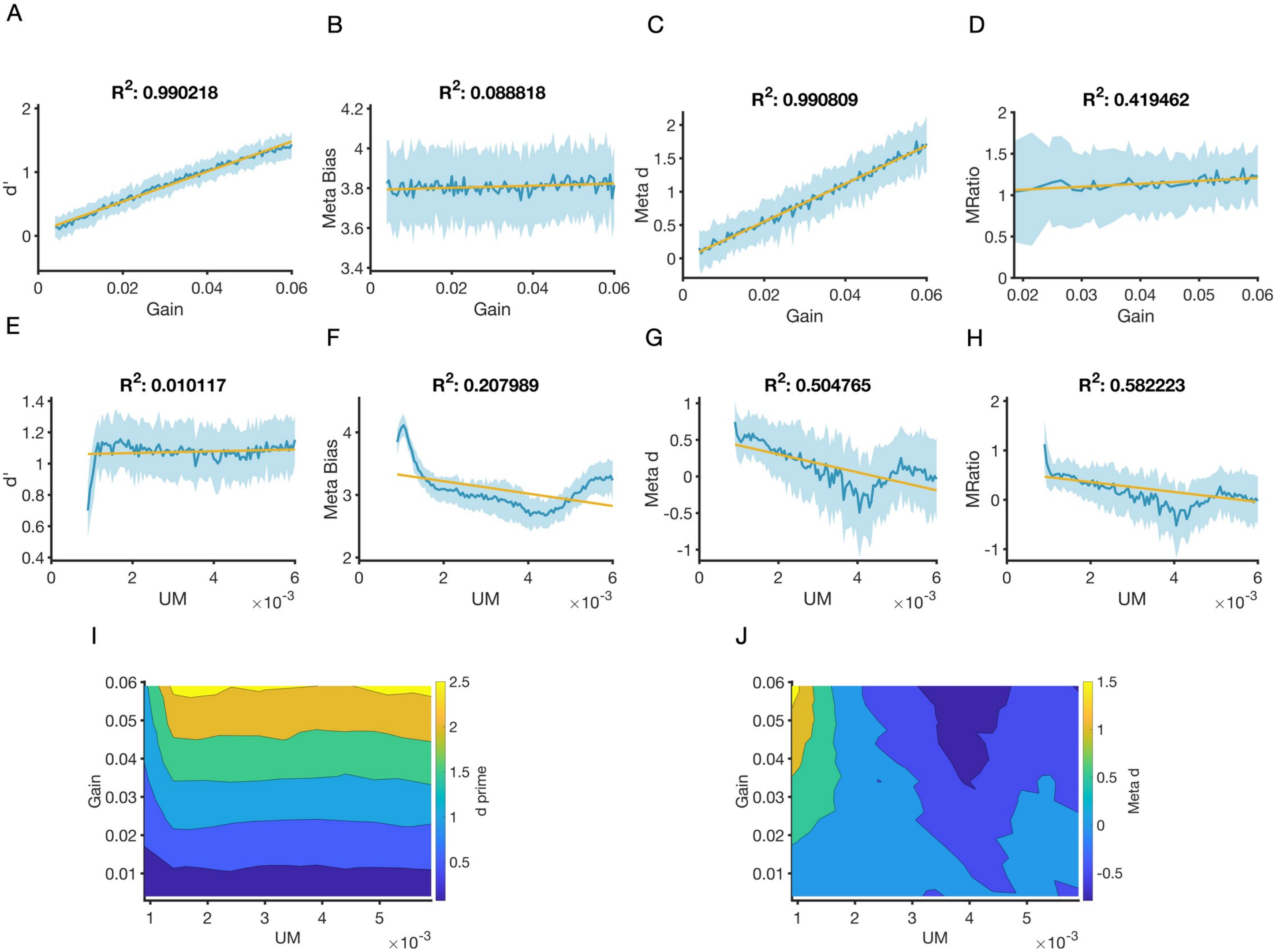

Fig 2. Dissociable changes in metacognition are associated with changes in uncertainty modulation. The behaviour of the model was analysed using standard metrics of performance (d') and metacognition (metacognitive bias, sensitivity (meta_d') and efficiency (meta_d'/d')). Blue line represents mean value of metric across 50 simulations. Shaded area is standard deviation. Yellow line is linear fit to mean value of metric as a function of parameter value. Increases in gain lead to monotonic increases in (**A**) d' ($\beta_1 = 0.5$, $R^2 = 0.99$, $p<0.001$) and (**C**) metacognitive sensitivity ($\beta_1 = 28.45$, $R^2 = 0.99$, $p<0.001$) but (**B**) a small effect on bias ($\beta_1 = 23.5$, $R^2 = 0.09$, $p<0.001$). Gain has a moderate positive weak negative effect on (**D**) metacognitive efficiency ($\beta_1 = 3.53$, $R^2 = 0.41$, $p<0.001$), possibly driven by the strong linear increase in d' in panel A. Increasing UM has no effect on (**E**) d' ($\beta_1 = 5.65$, $R^2 = 0.01$, $p = 0.15$), but a negative effect on (**F**) metacognitive bias ($\beta_1 = -122.83$, $R^2 = 0.2$, $p<0.001$), (**G**) metacognitive sensitivity ($\beta_1 = -98.83$, $R^2 = 0.5$, $p<0.001$), and (**H**) metacognitive efficiency ($\beta_1 = -100.49$, $R^2 = 0.58$, $p<0.001$). In (**I**-**J**), we varied both parameters and measured the effect on (**I**) d' and (**J**) metacognitive sensitivity. The increase in d' is mostly driven by changes in gain (**I**), whereas changes in metacognitive sensitivity are mostly driven by UM (**J**). All simulations were done with the same fixed list of dot differences (2.8 in log-space). In simulations (**A**-**H**), where the gain (UM) parameter is varied, UM (gain) was fixed at 0.0009 (0.0029). R^2 in all panels is adjusted R^2. Confidence data was generated by binning the uncertainty values into 6 bins, assuming equal bin width (see Methods). See Fig D in S5 Appendix for additional simulations with different parameter values. See also S5 Appendix for results that highlight dissociable changes in metacognitive bias as a result of varying UM.

https://doi.org/10.1371/journal.pcbi.1009201.g002

efficiency (Fig 2H). We then varied both parameters together and confirmed that changes in first-order task performance (d') (Fig 2I) are driven by changes in gain, whereas changes in metacognitive sensitivity ($meta{-}d'$) (Fig 2J) are driven by changes in UM.

The bottom row of Fig 2 suggests a linear fit is not sufficient to account for the relationship between UM and *d'*, meta bias, *meta_d'*, and *meta_d'/d'*. More specifically, despite *d'* remaining mostly constant (~1–1.1) in the majority of the explored UM parameter space (Fig 2E), Fig 2F shows that when UM is between 0.001 and 0.0015, *d'* increases when increasing UM. Furthermore, Fig 2F–2H show some inflection points where the behaviour before and after these points is different. For instance–meta bias peaks when UM = 0.0015, whereas *meta_d'* and *meta_d'/d'* peak at UM = 0.005. The existence of nonlinear relationships between UM and our theoretical measures of metacognition is perhaps not surprising given that the value of UM governs how two highly non-linear subnetworks of our model interact to generate decision performance and confidence (via increasing/decreasing excitatory feedback).

We note that the model's uncertainty does not in itself discriminate between correct and incorrect responses. In the model, such differences in confidence naturally emerge through the differences in response times for correct/incorrect trials as a function of difficulty. More specifically, giving the uncertainty-monitoring population more time to integrate input naturally leads to higher uncertainty (less confidence), which in turn is more likely to occur both during incorrect trials and on more difficult problems.

Overall, the results suggest that, in our model, a dissociable uncertainty-monitoring mechanism can drive changes in metacognition, in the absence of any change in task performance. More specifically, stronger uncertainty modulation is associated with a decrease in metacognitive sensitivity, bias, and efficiency, but not perceptual sensitivity. Armed with this understanding of how model parameters relate to facets of metacognitive performance, we next fit the model to subjects' data, and apply a computational psychiatry approach in order to relate variation in model parameters to psychopathology.

Model fits to subject data

We re-analysed data from [32], in which subjects (experiment 1: 498 subjects, experiment 2: 497 subjects) completed an online task via Amazon Mechanical Turk. In the task, upon initiating a trial, a fixation cross appears for 1000ms, followed by two black boxes each filled with a number of white dots (see Fig 1A). Subjects indicated first which box contains the greater number of dots, by pressing the right or left arrow key on a computer keyboard, and then provided their confidence rating on a numerical scale (1–11 for experiment 1, 1–6 for experiment 2).

To provide insight into the interaction between decision formation and metacognitive processes in this task, we simulated and fitted our neural circuit model of decision uncertainty to subjects' choices and response times [39,40]. This allowed us to use subjects' explicit confidence reports as an out-of-sample test of the model's ability to account for individual differences in metacognition. For simplicity, we only simulated the sensorimotor and uncertainty modules of the circuit, as originally introduced in [39,40] (See Fig 1B).

In fitting our model to subjects' choices and response times, we used a procedure based on the subplex optimisation method [52,53] (see Methods). The subplex optimisation method is an evolution of the simplex method [54]–one that is better suited for optimising noisy objective functions. Importantly, when parameterising our model, we initially set the values of all parameters to those found in our previous work [40], allowing only two parameters to vary in the fitting procedure. The first parameter is a 'gain' parameter, which maps the dot difference to input current flowing into the sensorimotor populations (see Methods). Subjects having larger values for the gain parameter generally have better choice accuracy, i.e., at the circuit level, a larger gain value implies a larger bias in sensory input to the sensorimotor population corresponding to the correct choice. The second parameter is the strength of uncertainty modulation (see Fig 1D for an example of effect of varying this parameter on the decision process).

In experiment 1, subjects completed a perceptual decision-making task in which they judged which box contained a greater number of dots, followed by a confidence report on an 11-point numerical scale. Subjects then completed a number of questionnaires to assess self-reported psychiatric symptoms (see Methods). Unsurprisingly, subjects were more accurate when the task was easy, i.e. when the difference between the number of dots was large (see Fig 3A). The model captures this straightforward relationship between accuracy and task difficulty (Fig 3A), and accounts for individual variation in accuracy levels (Fig 3C).

In line with existing findings from both human and animal studies of decision-making [46,55,56], subjects' correct (error) responses were quicker (slower) as the task became easier,

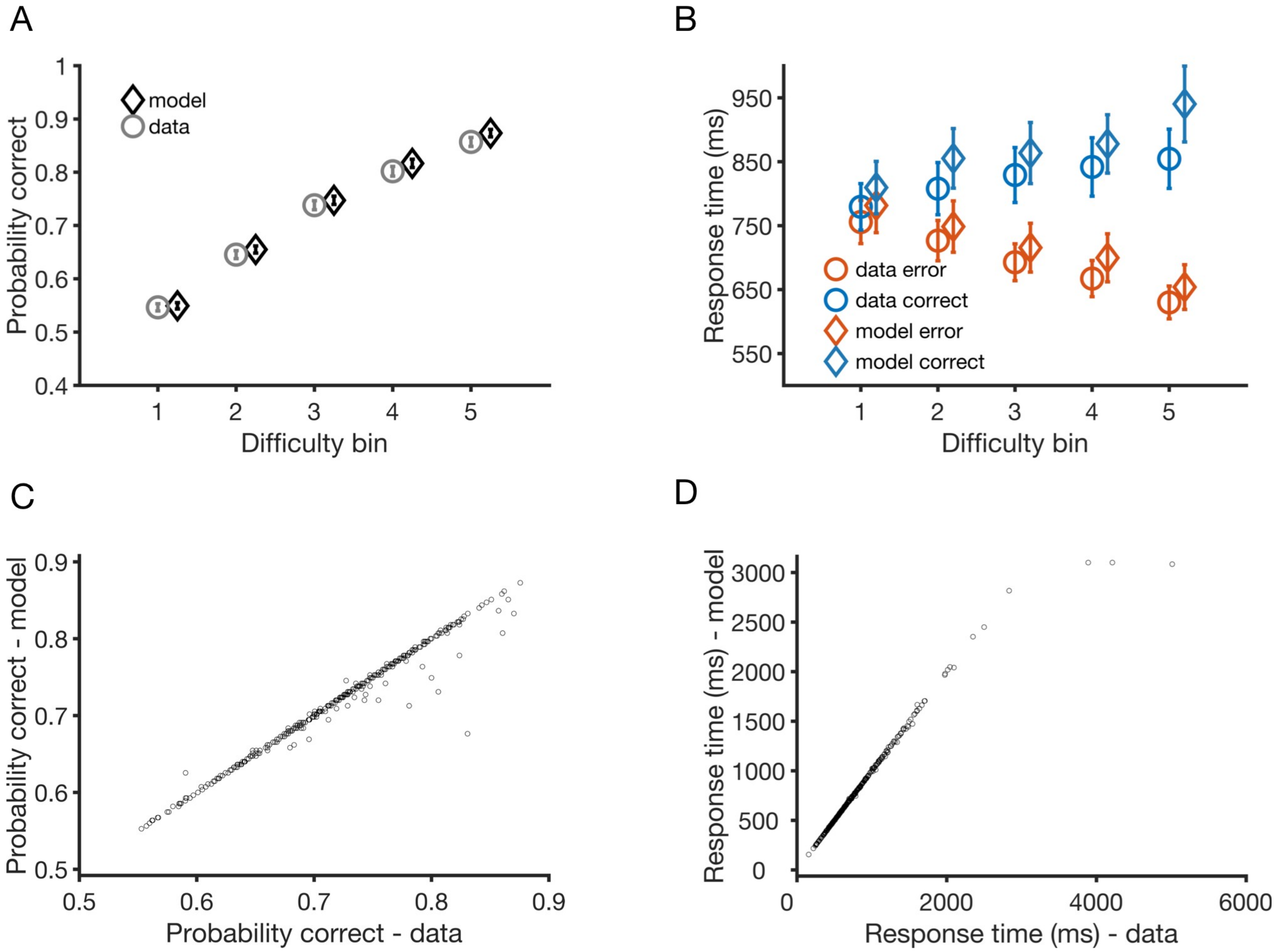

Fig 3. Model accounts for subjects' perceptual performance in experiment 1. A. Choice accuracy, i.e. probability correct as a function of task difficulty from experiment 1 of [32] averaged across all 498 participants. Task difficulty is split into 5 difficulty bins (1: most difficult, 5: easiest) as in the original paper (see Methods). Grey markers: data. Black markers: model fits. **B.** Response times as a function of task difficulty from the data (circles) and model fits (diamonds) averaged across all participants. Orange (blue) markers: Error (correct) responses. The typical '<' pattern, i.e. response times for correct (error) responses increasing (decreasing) as a function of task difficulty, is found in both the model and data. **C.** Scatter plot of observed (empirical) vs. simulated overall accuracy and **D.** response times for each of the 498 subjects. Error bars indicate 95% confidence interval. Random seed is reset after each simulation during the fitting procedure and for the purposes of generating panels **C** and **D** (but not **A** and **B**). See Fig C in S5 Appendix for scatter plots without resetting the random generator seed.

https://doi.org/10.1371/journal.pcbi.1009201.g003

forming a '<' pattern of response times as a function of difficulty (see Fig 3B and 3D for individual variation in mean response time). Observing an interaction between difficulty and accuracy in response time data is particularly striking given that the task was administered using a web-based platform, where response time measurement might be expected to be noisier than in standard laboratory settings. However, such a pattern was closely mirrored by our model fits, and importantly allowed us to constrain the model's estimates of subjects' confidence (see below).

Neuronal model constrained with perceptual performance accounts for subjects' confidence reports

We next asked whether our fitted model parameters could account for subjects' explicit confidence reports, even though these data had not been used to constrain the model. Here, we leverage the close relationship between confidence, response time and task difficulty to make inferences about trial-by-trial uncertainty (or confidence) levels from model fits to first-order performance [41,57]. In our model, longer response times allow more time for the uncertainty monitoring population to activate—leading to higher uncertainty (see Methods).

We first simulated our neural circuit model with the parameters fitted to subjects' choices and response times from experiment 1. We then applied distribution matching [46] to map the model's simulated uncertainty levels onto subjects' retrospective confidence reports. More specifically, instead of equal-width binning used in our analyses thus far, the shape of the overall mapping (i.e. prior to conditioning on performance or difficulty) is inferred from the distribution of experimental confidence reports, per subject (see Methods). This allowed us to show the model accounts for the complex relationship between decision confidence and task difficulty (see Fig 4A). The results also hold after conditioning confidence reports on trial outcome (i.e. correct vs. error). Importantly, these effects result from the intrinsic nonlinear dynamics of the network after fitting to (and constraining the model with) subjects' first-order performance data alone. The empirical confidence data are only used to set confidence thresholds, prior to conditioning on stimulus difficulty and accuracy. Hence the model is able to account for individual differences in subjects' perceptual and metacognitive performance despite model fits only having access to choices, response times and the overall distribution of confidence ratings. We next asked whether the parameters of in the model might also covary with psychiatric symptom scores.

Psychiatric symptoms are associated with the strength of uncertainty-monitoring

In experiment 1, upon completion of the main perceptual task, participants completed a series of standard self-report questionnaires that assess a range of psychiatric symptoms [58–67]. The questionnaires comprised: Zung Self-Rating Depression Scale, Generalized Anxiety Disorder 7-item scale, Short Scales for Measuring Schizotypy, Barratt Impulsiveness Scale 11, Obsessive-Compulsive Inventory-Revised [OCI-R], and Liebowitz Social Anxiety Scale.

As in [32], we ran a series of linear regressions to tease apart the relationship between psychiatric symptoms and model parameters. Importantly, here, we were able to account for differences in perceptual and metacognitive performance using only two model parameters, as highlighted in our model fits above. The first parameter (UM) controls the strength of uncertainty modulation. The second (gain) parameter maps the dot difference subjects see on the screen to difference in input current flowing into the model's sensorimotor neuronal populations.

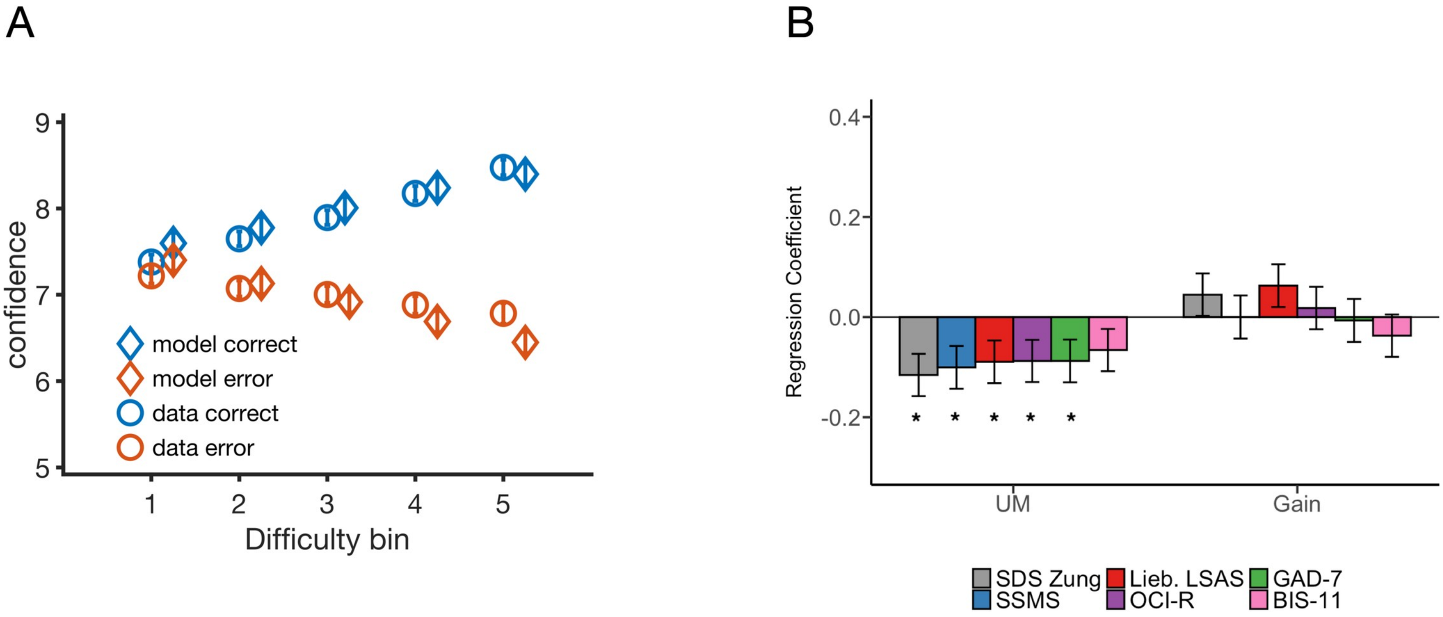

Fig 4. Model accounts for subjects' confidence reports and individual differences in uncertainty modulation predict symptom scores. A. Confidence reports averaged across all participants from experiment 1 data (circles) and model (diamond) as a function of task difficulty. Orange (blue) markers: Error (correct) responses. Note that the model was fit only to first-order performance data (accuracy and response times) and fits to confidence represent an out-of-sample prediction. Confidence increases (decreases) as a function of changing task difficulty for correct (error) responses. **B.** Symptom scores from experiment 1 were entered into a multiple regression model predicting the strength of uncertainty modulation and gain parameters from the model fits to task performance (choices and response times). Self-report measures of depression (grey), schizotopy (blue), social anxiety (red), obsessive and compulsive symptoms (purple) and generalised anxiety (green) are significantly associated with weaker uncertainty modulation. No significant association was found between impulsivity (pink) and the strength of uncertainty modulation. No significant association was found between the symptom scores and the gain parameter. See Methods for details on the regression models. Error bars indicate s.e.m. All regression results shown control for the influence of age, gender, and IQ (see Fig B in S5 Appendix for regression model results with age and IQ predicting model parameters). * p<0.05.

https://doi.org/10.1371/journal.pcbi.1009201.g004

We entered each questionnaire score (see Methods) into multiple linear regressions predicting the uncertainty modulation and gain parameters. The results (see Fig 4B) show that increases in z-scored self-reported scores were broadly associated with weaker uncertainty modulation across all dimensions of psychopathology, with the exception of impulsivity, though the association strengths did not differ between questionnaires. This contrasts with the gain parameter, which did not correlate with any of the self-reported scores (p>0.05) in experiment 1. These results largely recapitulate the relationships between empirical confidence level and psychiatric symptoms scores (albeit with minor differences in effect sizes) observed in [32], but now provide a potential circuit-level explanation for such differences (i.e., a change in the strength of uncertainty modulation).

We also followed the same approach for experiment 2 (see S2 Appendix), although here we found no significant association between the majority of self-reported scores (or cross-cutting factors derived from these scores, see Fig A in S2 Appendix) and model parameters. This lack of significance in experiment 2 may reflect the smaller variance in difficulty (due to the staircase procedure) leading to inferences on uncertainty modulation being less constrained by the data (see Fig C in S2 Appendix). To explore this further, we attempted to recover the parameters fitted to both experiment 1 and 2 data and found that the fits to experiment 1 data were indeed more stable–potentially due to the larger variation in task difficulty. We note however that qualitatively, similar symptom scores (e.g. depression, anxiety) that were negatively related to uncertainty modulation in experiment 1 were also negatively related to the uncertainty modulation in experiment 2. In addition, when using the HMeta-d toolbox [21] to perform a hierarchical regression [68], we obtained a positive association between the strength of uncertainty modulation and metacognitive efficiency in experiment 2 (mean value of μ_β = 0.0516, 95% highest density interval = (0.0813, 0.0016), see Fig A in S5 Appendix).

Discussion

While self-reported psychiatric symptoms have been shown to be associated with dissociable differences in metacognition, the mechanisms underlying such changes have remained elusive. In this work, using a computational circuit model of decision-making, we show that shifts in metacognition are associated with disturbances in the interaction between decision-making and uncertainty-monitoring networks. Specifically, uncertainty modulation is associated with metacognitive bias, sensitivity, and efficiency. Importantly, changes in uncertainty modulation strength have no effect on perceptual sensitivity. Notably, our model-fitting approach enabled inferences about uncertainty modulation (and, in turn, these facets of metacognition) from fits to first-order performance data alone. The empirical confidence distributions were used to set confidence thresholds alone (via distribution matching), thus influencing the overall mean and shape of the modelled confidence distribution, but not the relation between confidence and features of performance, response time or task difficulty. Nevertheless, the model is able to account for individual differences in subjects' perceptual and metacognitive performance despite model parameters being adjusted solely on the basis of patterns of choices and response times. When we apply this approach to data from an online perceptual decision task, we find that self-reported psychiatric symptoms are associated with disturbances in uncertainty modulation.

Through a dedicated uncertainty-monitoring population, our model of decision uncertainty captures key features of the neurobiology of metacognition, while remaining sufficiently simple to fit to data. Recent work has shown that long response times are associated with lower confidence for an impending decision [40,41,57]. Our computational model naturally accounts for this phenomenon. More specifically, winner-take all behaviour is less prevalent when the external stimulus input to the network is (or close to) symmetric, i.e. when stimulus information is ambiguous. This high level of competition between the sensorimotor populations prolongs the time taken to reach a decision threshold, and by allowing more time for an uncertainty-monitoring module to integrate bottom-up input results in higher uncertainty. Building on this proposed mechanism, and existing behavioural evidence, our approach allows us to infer metacognitive performance from first-order (i.e. response time) data.

Crucially, we go beyond simply relating our model dynamics to decision confidence [39]. By analysing our model's uncertainty estimates using standard metrics of metacognition, we reveal that changes in uncertainty modulation in such a network has effects on metacognitive bias, sensitivity, and efficiency, while leaving perceptual sensitivity unaffected. In simulation, there were nonlinear relationships between UM and meta-d, with both increases and decreases affecting metacognitive bias and sensitivity. In contrast, in fits to empirical data, decreases in UM were generally associated with greater psychopathology and lower metacognitive sensitivity. It is of interest to note here that such dissociable changes in metacognitive ability, as a result of a (higher-order) disturbance in the strength of uncertainty modulation, finds support in recent neuropsychological work. For instance, lesions in prefrontal brain regions are associated with deficits in metacognitive ability, but not task performance [28,29,31], highlighting the contribution of higher-order brain regions to metacognition [26,27]. Future work could combine our computational framework with neuroimaging to further elucidate the neural basis of metacognitive ability.

Our model architecture complements work exploring how parallel sensorimotor neural populations, with different normalisation tuning, may account for cases where confidence is altered in the absence of a performance change [44]. We anticipate that a full circuit model of metacognition will need to combine aspects of parallel evidence accumulation (simulating e.g. parietal cortex neural populations) and higher-order monitoring (simulating e.g. prefrontal neural populations involved in the representation and use of uncertainty [30]). Of particular note here is that we show changes in metacognitive bias may also occur due to shifts in parameters governing higher-order nodes of the circuit. It seems plausible that different forms of confidence bias could map onto different levels of the system. For instance, confidence shifts induced by changes in volatility or amount of evidence may be explained by preferential activation of less-normalisation-tuned populations [44], whereas confidence biases related to mood or more "global" aspects of performance may be explained by changes in higher-order nodes of the system [69].

Adopting a computational psychiatry approach, we shed light on a potential driver of metacognitive distortions reported in recent work in relation to mental health symptoms [32]. Rouault and colleagues [32] showed that symptom scores for depression, social anxiety, and generalised anxiety relate to lower confidence level. In the present report, following similar analyses, we show that these relationships can be explained by changes in the strength of uncertainty modulation, in the absence of any change in sensory gain. Our analyses not only recapitulate previously-reported relationships with depression and anxiety (Fig 5), but show that schizotopy and OCD scores also relate to disturbances in uncertainty modulation [70], in line with existing work relating deficits in self-evaluation to schizophrenia [71].

In Fig 5, we compare symptom scores predicting UM (on the left in Fig 5 below), symptom scores predicting empirical mean confidence (Fig 5, middle), and symptom scores predicting metacognitive sensitivity (Fig 5, right). Fig 5 shows that the standardised effect sizes are slightly

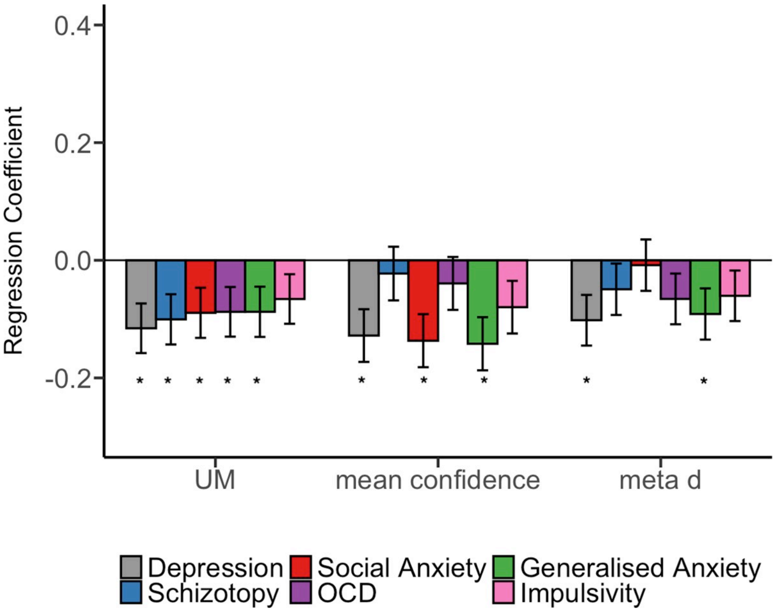

Fig 5. UM parameter offers an implicit, low-dimensional marker of metacognitive (dys)function. Symptom scores from experiment 1 were entered into a multiple regression model predicting the strength of uncertainty modulation, empirical mean confidence, and metacognitive sensitivity. Self-report measures of depression (grey), schizotopy (blue), social anxiety (red), obsessive and compulsive symptoms (purple) and generalised anxiety (green) are significantly associated with weaker uncertainty modulation. Depression, social anxiety, and generalised anxiety are associated with lower mean confidence. Depression and generalised anxiety are associated with decreased metacognitive sensitivity. * $p<0.05$.

https://doi.org/10.1371/journal.pcbi.1009201.g005

larger in the case of Depression (β = -0.128, $p<0.05$), Social Anxiety (β = -0.136, $p<0.05$), and Generalised Anxiety (β = -0.141, $p<0.05$) predicting mean confidence, compared to predicting UM (β = -0.115, $p<0.05$ in the case of Depression, β = -0.089, $p<0.05$ in the case of Social Anxiety, and β = -0.087, $p<0.05$ in the case of Generalised Anxiety). The difference is smaller when comparing the standardised effect sizes in the case of symptom scores predicting UM vs. metacognitive sensitivity (β = -0.102, $p<0.05$ in the case of Depression, and β = -0.091, $p<0.05$ in the case of Generalised Anxiety). Overall, the results suggest that the UM parameter offers an implicit, low-dimensional marker of metacognitive (dys)function, but that confidence rating data would still give a richer perspective and provide distinct measures of sensitivity and bias (as shown in our Fig A in S2 Appendix).

Recent work has demonstrated that symptoms of OCD are associated with deficits in utilising evidence to update confidence [35]. In the context of our model, this can be explained by the weaker UM strength associated with Obsessive-Compulsive Inventory–Revised (OCIR) scores—i.e. participants with higher OCIR scores tend to monitor uncertainty for longer, prolonging their response times, but not necessarily increasing their confidence in their decisions. Such a mechanism is supported by recent work linking extended evidence accumulation associated with compulsive behaviour to increased decision-making thresholds and metacognitive impairments [72,73]. Notably, in the current work, we could account for individual differences in task (Figs 3 and 4) and metacognitive performance (Fig 4A) even in large samples of data (N = 495 in Experiment 1, N = 496 in Experiment 2 –see S1 and S2 Appendices for Experiment 2 results) collected over the web where experimental control over subjects' responses is less precise, and response time measurement potentially noisier. Taken together, the results from both experiments suggest our computational framework can be used to study the interaction between metacognition and psychiatric symptoms without requiring subjects to explicitly report confidence in decisions—potentially opening the door to using shorter, more engaging tasks such as smartphone games [74].

We also explored whether our model accounts for metacognition-psychopathology relationships in a task with staircased difficulty levels (experiment 2 in [32]). Although our analyses of the UM parameter show a similar pattern to those obtained for metacognitive bias in the original study (Fig A in S2 Appendix), these relationships between factor scores and model parameters did not reach significance. One interpretation of this equivocal result is that effective inference on individual differences in uncertainty modulation strength may require perceptual tasks with systematic variation in difficulty, to enable full coverage of the RT-accuracy-difficulty surface (i.e. the < patterns). Importantly, we found that the fit for experiment 2 is not as stable as the fit for experiment 1 (Figs B and C in S2 Appendix). Further theoretical work is needed to determine the effect of per-subject difficulty variance on the ability to infer such model parameters.

There are notable limitations to the scope of the model that deserve further investigation in future work. First, it is worth noting that our previous work [39] showed that our model may produce uncertainty and response time patterns that do not strictly follow the '<' pattern, e.g. with less pronounced increase (decrease) in response times (confidence) for incorrect trials. Such patterns have been previously reported in empirical data [57,75–77] when considering a wider range of task types (e.g. free-response tasks), and stimulus durations. Second, in our model, there exists a high positive correlation between the maximum firing rate achieved during the trial for both the losing sensorimotor population and the model's uncertainty monitoring population. This mechanism may limit the model's ability to account for settings where high (low) confidence is associated with slow (fast) response times. Future modelling work can investigate fitting the model to data from such settings. Finally, previous work [45] has shown that slower response times can be associated with higher confidence when subjects optimise for accuracy over speed. In our model, both the non-selective top-down excitation of the sensorimotor populations and the decision time (which determines the integration window of the uncertainty-monitoring population) contribute to a high positive correlation between uncertainty and response time. Accounting for a reversal in this relationship is beyond the scope of the current modelling work. In order to account for various speed/accuracy trade-offs leading to differential confidence-reponse time correlations, future work will need to investigate alternative model architectures where the integration-time window is fixed across trials, and is not decision-time dependent.

Previous versions of our neural circuit model have also been applied to tasks with explicit motor reaching trajectories through a dedicated motor output network [39,40]. Here, given that participants reported their decisions using a keyboard button press rather than continuous motor responses, this aspect of the network was less relevant. However, our current findings highlight the promise of leveraging the full model to dissect the interaction between uncertainty-monitoring, indecisiveness and psychiatric symptoms in a task where both sensory input and motor output are quantified in a continuous, dynamic fashion. Because these relationships can be obtained from fits to first-order performance and response time data alone, future work could leverage our computational framework to infer facets of metacognition in situations where obtaining explicit metacognitive judgements is problematic or impossible, e.g. in studies of animals or children.

In summary, we employed a biologically-plausible model of decision uncertainty to relate dissociable shifts in metacognition to isolated disturbances in uncertainty modulation. We validate our model against empirical data, and relate its parameters to psychopathology. Our work bridges a gap between a biologically plausible model of confidence formation and the observed disturbances in metacognition seen in mental health disorders, and provides a first step towards mapping theoretical constructs of metacognition onto dynamical models of decision uncertainty. In doing so, we provide a computational framework for modelling metacognitive performance in settings where access to explicit confidence reports is either difficult or impossible.

Methods

Ethics statement

Data analysed in this work was first collected as part of a study conducted by [32]. Participants provided written consent in accordance with procedures approved by the University College London Research Ethics Committee (Project ID 1260/003).

Neural circuit model of uncertainty

We modelled the processes underpinning decisions and confidence using a neural circuit model of uncertainty described previously [39,40]. The version of the model used here comprises two interacting subnetworks—a decision-making *sensorimotor module*, and an *uncertainty-monitoring* population.

As in previous work [39,40], the sensorimotor module is modelled using a reduced (i.e. two-variable) spiking neural network model [36,38]. The dynamics of the neuronal populations are described by:

$$\frac{dS_L}{dt} = -\frac{S_L}{\tau_s} + (1 - S_L)\gamma H(x_L, x_R) \tag{1}$$

$$\frac{dS_R}{dt} = -\frac{S_R}{\tau_s} + (1 - S_R)\gamma H(x_R, x_L) \tag{2}$$

where S_L and S_R are the synaptic gating variables for the sensorimotor population selective to leftward and rightward stimulus information, respectively. τ_s denotes the synaptic gating time constant. γ is a constant that is derived in previous theoretical work [36] that describes a reduction of the original spiking neuronal network model of decision making [38].

The firing rate of a sensorimotor population can be described using the nonlinear function H:

$$H_i = \frac{ax_i - b}{1 - e^{-d(ax_i - b)}} \tag{3}$$

where a, b, d are parameters fitted to the leaky integrate-and-fire model [38]. The variable i can be L or R, denoting sensorimotor population selective for rightward or leftward sensory information, respectively. x_i denotes the total input into population i, and can be described by:

$$x_i = w_+ S_i - w_- S_j + I_c + I_i + I_\sigma + w_u U \tag{4}$$

where w_+ denotes synaptic weight for self-excitation, whereas w_- denotes synaptic weight for mutual inhibition. I_c is some constant input. I_σ denotes noise—here we use the same noise described by an Ornstein–-Uhlenbeck process as in [36]. I_i denotes external input flowing into population i, as a function of the dot difference participants see on the screen (Fig 1). This external input is described by:

$$I_i = w_e \mu_0 (1 \pm \varepsilon) \tag{5}$$

where w_e is a synaptic weight, whereas μ_0 is some baseline external input. ε can be described by:

$$\varepsilon = \text{gain} \cdot \text{dot difference} \tag{6}$$

where the input gain parameter maps the dot difference to difference in input flowing into the sensorimotor populations. In our model, sensorimotor populations continue to integrate evidence for 180ms after the initial decision is made (nondecision time). This nondecision time has been used in previous work to account for signal transduction delays [78,79]. Our model does not account for more extended post-decisional processing, or the incorporation of new, post-decisional evidence. Such additions to the model may usefully augment the extent to which we are able to accommodate dissociations between confidence and performance.

Importantly, the last term in Eq (4) ($w_u U$) determines the strength of feedback excitation from the uncertainty-monitoring neuronal population. More specifically, w_u is referred to throughout this article as UM, or uncertainty modulation strength. U denotes the dynamical variable of the uncertainty-monitoring population, which is described by:

$$\tau_u \frac{dU}{dt} = [H_L + H_R - l]_+ - U \tag{7}$$

where $[]_+$ is a threshold linear function (threshold = 0). H_L and H_R are functions denoting firing rates for sensorimotor populations selective for leftward and rightward stimulus information, respectively (from Eqs (1) and (2)). l denotes some constant input that suppresses the firing of the uncertainty-monitoring population. This input is de-activated 200ms after stimulus onset, and is reactivated when one the firing rate of the sensorimotor populations reaches a decision threshold (see Fig 1). Eq (7) includes a leak term ($-U$), hence why the integration decays over time when no external input is present (i.e., a leaky integrator). We summarise the values of all model parameters in Table 1.

Table 1. Table of fixed model parameter values for all participants. Parameters τ_S, τ_u, a, b, d, I_c, w_e were directly adapted from [40]. Parameters μ_0, w_+, S_{th} were manually tuned to adapt the model simulations to the task and stimuli.

Parameter	Description	Value
τ_S	Synaptic gating time constant	100ms
τ_u	Uncertainty population time constant	150 ms
a	Input-output function parameter	270 (V nC)$^{-1}$
b	Input-output function parameter	108 Hz
d	Input-output function parameter	0.154 s
I_c	External tonic input	0.3255 nA
w_+	Self-excitation strength	0.261 nA
w_-	Inhibition strength	0.0497 nA
μ_0	Baseline stimulus input	26.49 Hz
w_e	External input synaptic strength	0.00052 nA Hz^{-1}

https://doi.org/10.1371/journal.pcbi.1009201.t001

Quantifying uncertainty within a trial

As in our previous work [40], for a given trial, we used the maximum firing rate value of the uncertainty-monitoring neuronal population as a decision uncertainty measurement for that particular trial (the inverse of decision confidence). When extrapolating confidence reports from simulations (e.g. for Fig 2 simulations), we used simple equal-width binning in 6 bins to relate continuous uncertainty measurements to a 6-point confidence scale, similar to the one used in experiment 2.

Each participant uses the confidence scale differently, e.g. on a 6-point probabilistic scale, one might consistently pick 5 as their highest confidence level. In order to relate simulated uncertainty to empirical confidence data from each participant, we match the distribution of simulated uncertainty to the marginal distribution of empirical confidence reports (i.e. prior to conditioning on accuracy, response times, or difficulty [46]). More specifically, per subject, we (non-parametrically) infer the shape of the mapping from their experimental confidence distribution. First, we compute the cumulative distribution function (CDF) of their full confidence distribution. Then, we use this CDF to derive binning width thresholds. The thresholds here represent the quantiles of the subjects' simulated confidence for the probabilities represented by CDF computed from experimental confidence distribution.

Model fitting procedure

To fit our model to participants' first order performance, we used a procedure that exploits the subplex optimisation method [52,53]. Subplex optimisation is based on the simplex optimsation method, but adapted for noisy objective functions [53]. For each participant, we minimise the cost function:

$$\text{cost} = \frac{1}{m}(RT_{\text{model}} - RT_{\text{data}})^2 + \frac{1}{n}(\text{accuracy}_{\text{model}} - \text{accuracy}_{\text{data}})^2 \tag{8}$$

where RT_{model} is the model's mean response time from a single model simulation (with a fixed random seed), RT_{data} denotes the participants' mean response time. Similarly, $\text{accuracy}_{\text{model}}$ and $\text{accuracy}_{\text{data}}$ denote overall accuracy for the model and experiment, respectively. m and n are normalisation terms for response times and accuracy, respectively. Here, m and n are set to the model statistic (i.e., $m = \text{RT}_{\text{model}}$, and $n = \text{accuracy}_{\text{model}}$) [52]. The cost function can be calculated per difficulty level (see [80]). Here, we opted for calculating the cost using the overall accuracy and overall response times (across all difficulties). Importantly, we only fit two free parameters: gain and w_u, from Eqs (6) and (4), respectively. The vast majority of the other model parameters are adapted from our previous work [40] (see Table 1). When generating synthetic data using the model (for fitting or otherwise), for experiment 1, we simulate 210 trials while generating dot difference data from a uniform distribution bounded by the max and min value for each difficulty block as found in the data. For experiment 2, we simulate the model with the vector of dot differences experienced by each participant.

Participants

We re-analysed data from [32], and the reader is referred to this paper for a full description of the task and sample. All participants were recruited over the web using Amazon Mechanical Turk. In experiment 1, 663 (498 after exclusions) participants completed the task, and were 18–75 years of age. In experiment 2, 637 (497 after exclusions) participants completed the task, and were 18–70 years of age. The study protocol was approved by the University College London Research Ethics Committee (REF 1260/003) and all participants provided informed consent before undertaking the task. All participants in experiment 1 and 2 were compensated $4. A $2 bonus was paid out to participants on two conditions: In experiment 1, the bonus was paid if participants achieved >50% accuracy in task performance, and passed a check question. In experiment 2, the bonus was paid if participants achieved task performance between 60–85%, and passed a check question. We used the same exclusion criteria applied in [32] and described in the Supplementary Material of that paper.

Task

In both experiments, participants completed a simple perceptual decision-making task where they judged (using a keyboard press) which box contained a higher number of dots, with no feedback. One box was always half-filled (313 dots out of 625 positions), while the other box contained an increment of +1 to +70 dots compared to the standard. In any given trial, a fixation cross first appeared for 1 second, followed by two black boxes with two different amounts of dots (for 300ms). The position of the box with higher number of dots (i.e. target box) was pseudo-randomised. After indicating the position of the target box (left/right) via a keyboard arrow button press, the box was highlighted for 500ms. In experiment 1, participants completed 210 trials, split over 5 blocks, where the difficulty was varied. The position of the target box was pseudo-randomised across all trials and within each of 5 difficulty bins.

After every trial, participants provided a confidence judgement on a full 11-point probabilistic scale: 1 = certainly wrong, 3 = probably wrong, 5 = maybe wrong, 7 = maybe correct, 9 = probably correct, 11 = certainly correct. Finally, pre- and post-task global confidence ratings were given by participants, together with their estimates of expected maximum and minimum levels of task performance.

Prior to undertaking the experiment, participants were required to select on an 11-point scale their global expected performance level in the task relative to others, together with a maximum and minimum expected performance level. After completing the task, participants were again asked to rate their expected performance level in the task relative to others, using the same scale. Pre- and post- global confidence levels were not analysed here.

Experiment 2 (see S1 Appendix) is identical to experiment 1 in all but three aspects. First, [32] used a staircase (calibration) procedure to fix participants' perceptual performance [26,81]. The staircase procedure was two-down one-up, with equal step sizes. Step-sizes (in logspace) were: 0.4 for first 5 trials, 0.2 for next 5, 0.1 for the rest of the task. The starting point was 4.2. Each participant completed 25 practice trials at the beginning of the task to minimise the burn-in period. Second, participants reported their confidence on a 6-point confidence scale which ranged from 1 = guessing to 6 = certainly correct). Third, pre- and post-task global confidence ratings were omitted from experiment 2.

The entire experiment was coded in JavaScript with JsPsych version 4.3 [82].

Psychiatric questionnaires

Participants completed a set of self-report questionnaires used to assess their psychiatric symptoms [32]. In experiment 1, the questionnaires were:

- Depression using the Self-Rating Depression Scale (SDS) [58]
- Generalised anxiety using the Generalised Anxiety Disorder 7-Item Scale (GAD-7) [59]
- Schizotypy using the Short Scales for Measuring Schizotypy (SSMS) [60]
- Impulsivity using the Barratt Impulsiveness Scale (BIS-11) [61]
- Obsessive Compulsive Disorder (OCD) using the Obsessive-Compulsive Inventory–Revised (OCI-R) [62]
- Social anxiety using the Liebowitz Social Anxiety Scale (LSAS) [63]

In experiment 2 (S1 and S2 Appendices), the following changes were made to the set of questionnaires:

- Generalised Anxiety questionnaire was replaced by the State Trait Anxiety Inventory (STAI) Form Y-2 [64]
- Alcoholism was assessed with the Alcohol Use Disorders Identification Test (AUDIT) [65]
- Apathy was assessed with the Apathy Evaluation Scale (AES) [66]
- Eating disorders was assessed with the Eating Attitudes Test (EAT-26) [67]

These changes in experiment 2 were made to facilitate identification of three latent factors that accounted for the majority of covariance across individual questionnaire items [83].

Factor analysis

For experiment 2 data (see S1 and S2 Appendices), we obtained three latent factors that explain the shared variance across the 209 questionnaire items. To do that, we followed the same approach in [32,83], and used the *fa()* function from the Psych package in R. The three latent factors were Anxious-Depression, Compulsive Behaviour and Intrusive Thought, and Social Withdrawal.

Linear regressions

To estimate the relationship between the neural model parameters and self-reported psychiatric scores, we followed the same approach as in [32]. All regressors were z-scored to ensure comparability of regression coefficients. For each symptom score, and controlling for age, IQ and gender the regressions were:

$$\text{Param} = \beta_0 + \beta_1\text{Score} + \beta_2\text{Age} + \beta_3\text{Gender} + \beta_4\text{IQ} \tag{9}$$

To assess the relationship between model parameters and the latent factor scores (see above), the regression was:

$$\text{Param} = \beta_0 + \beta_1\text{Factor 1} + \beta_2\text{Factor 2} + \beta_3\text{Factor 3} + \beta_4\text{Age} + \beta_5\text{Gender} + \beta_6\text{IQ} \tag{10}$$

Finally, we used linear regressions to estimate the contribution of two of the model parameters to standard metrics of metacognition and perceptual sensitivity. Here, we did not z-score the regressors as the goal was to visualise the relationship rather than quantitatively compare coefficients. The regressions were:

$$\text{metric} = \beta_0 + \beta_1 \text{model param} \tag{11}$$

Metacognitive bias, sensitivity, and efficiency

Metacognitive bias was computed as the mean confidence level across both correct and incorrect trials. To estimate metacognitive sensitivity, we entered simulated confidence reports as data in a Bayesian model of metacognitive efficiency, HMeta-d [21]. The model returns a value of metacognitive sensitivity ($meta{-}d'$) for each simulated dataset. To compute metacognitive efficiency, we calculated the ratio $meta{-}d'/d'$.

Supporting information

S1 Appendix. Model fit to experiment 2 data from Rouault et al. (2018).
(DOCX)

S2 Appendix. Relationships between psychiatric symptoms and model parameters in Experiment 2.
(DOCX)

S3 Appendix. Changes in metacognitive bias driven by UM.
(DOCX)

S4 Appendix. Experiment 1 fit with holdout.
(DOCX)

S5 Appendix. Other supporting figures.
(DOCX)

S1 Text. Exclusion Criteria.
(DOCX)

S2 Text. Integration onset timing parameter.
(DOCX)

S3 Text. UM as a low dimensional marker of metacognitive profile.
(DOCX)

Acknowledgments

We thank Marion Rouault for helpful discussions.

Author Contributions

Conceptualization: Nadim A. A. Atiya, Stephen M. Fleming.

Data curation: Nadim A. A. Atiya, Stephen M. Fleming.

Formal analysis: Nadim A. A. Atiya.

Funding acquisition: Raymond J. Dolan, Stephen M. Fleming.

Investigation: Nadim A. A. Atiya, Stephen M. Fleming.

Methodology: Nadim A. A. Atiya, Quentin J. M. Huys.

Project administration: Raymond J. Dolan, Stephen M. Fleming.

Resources: Raymond J. Dolan, Stephen M. Fleming.

Software: Nadim A. A. Atiya.

Supervision: Quentin J. M. Huys, Raymond J. Dolan, Stephen M. Fleming.

Validation: Nadim A. A. Atiya, Quentin J. M. Huys, Stephen M. Fleming.

Visualization: Nadim A. A. Atiya.

Writing – original draft: Nadim A. A. Atiya, Stephen M. Fleming.

Writing – review & editing: Nadim A. A. Atiya, Quentin J. M. Huys, Raymond J. Dolan, Stephen M. Fleming.

References

1. Friston KJ, Stephan KE, Montague R, Dolan RJ. Computational psychiatry: the brain as a phantastic organ. The Lancet Psychiatry. 2014; 1(2):148–58. https://doi.org/10.1016/S2215-0366(14)70275-5 PMID: 26360579
2. Huys QJM, Maia TV, Frank MJ. Computational psychiatry as a bridge from neuroscience to clinical applications. Nat Neurosci. 2016; 19(3):404–13. https://doi.org/10.1038/nn.4238 PMID: 26906507.
3. Wang X-J, Krystal JH. Computational psychiatry. Neuron. 2014; 84(3):638–54. Epub 2014/11/05. https://doi.org/10.1016/j.neuron.2014.10.018 PMID: 25442941.
4. Montague PR, Dolan RJ, Friston KJ, Dayan P. Computational psychiatry. Trends Cogn Sci. 2012; 16 (1):72–80. Epub 2011/12/14. https://doi.org/10.1016/j.tics.2011.11.018 PMID: 22177032.
5. Ratcliff R. A theory of memory retrieval. Psychological Review. 1978; 85(2):59–108. https://doi.org/10.1037/0033-295x.85.2.59
6. Ratcliff R, Smith PL, McKoon G. Modeling Regularities in Response Time and Accuracy Data with the Diffusion Model. Curr Dir Psychol Sci. 2015; 24(6):458–70. https://doi.org/10.1177/0963721415596228 PMID: 26722193.
7. Rescorla RA. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. Current research and theory. 1972:64–99.
8. Sutton RS, Barto AG. Reinforcement learning: An introduction: MIT press; 2018.
9. Marr D, Poggio T. From understanding computation to understanding neural circuitry. 1976.
10. Schultz W. The Reward Signal of Midbrain Dopamine Neurons. Physiology. 1999; 14(6):249–55. https://doi.org/10.1152/physiologyonline.1999.14.6.249 PMID: 11390860
11. Dayan P, Balleine BW. Reward, Motivation, and Reinforcement Learning. Neuron. 2002; 36(2):285–98. https://doi.org/10.1016/s0896-6273(02)00963-7 PMID: 12383782
12. Dolan RJ, Dayan P. Goals and habits in the brain. Neuron. 2013; 80(2):312–25. https://doi.org/10.1016/j.neuron.2013.09.007 PMID: 24139036.
13. Stephan KE, Bach DR, Fletcher PC, Flint J, Frank MJ, Friston KJ, et al. Charting the landscape of priority problems in psychiatry, part 1: classification and diagnosis. The Lancet Psychiatry. 2016; 3(1):77–83. https://doi.org/10.1016/S2215-0366(15)00361-2 PMID: 26573970
14. Murray JD, Anticevic A, Gancsos M, Ichinose M, Corlett PR, Krystal JH, et al. Linking microcircuit dysfunction to cognitive impairment: effects of disinhibition associated with schizophrenia in a cortical working memory model. Cereb Cortex. 2014; 24(4):859–72. Epub 2012/11/29. https://doi.org/10.1093/cercor/bhs370 PMID: 23203979.
15. Krystal JH, Murray JD, Chekroud AM, Corlett PR, Yang G, Wang X-J, et al. Computational Psychiatry and the Challenge of Schizophrenia. Schizophr Bull. 2017; 43(3):473–5. https://doi.org/10.1093/schbul/sbx025 PMID: 28338845.
16. Rolls ET, Loh M, Deco G. An attractor hypothesis of obsessive-compulsive disorder. European Journal of Neuroscience. 2008; 28(4):782–93. https://doi.org/10.1111/j.1460-9568.2008.06379.x PMID: 18671737
17. Dima D, Roiser JP, Dietrich DE, Bonnemann C, Lanfermann H, Emrich HM, et al. Understanding why patients with schizophrenia do not perceive the hollow-mask illusion using dynamic causal modelling. NeuroImage. 2009; 46(4):1180–6. https://doi.org/10.1016/j.neuroimage.2009.03.033 PMID: 19327402

18. Yang GJ, Murray JD, Repovs G, Cole MW, Savic A, Glasser MF, et al. Altered global brain signal in schizophrenia. Proc Natl Acad Sci U S A. 2014; 111(20):7438–43. Epub 2014/05/05. https://doi.org/10.1073/pnas.1405289111 PMID: 24799682.

19. Cohen JD, Servan-Schreiber D. Context, cortex, and dopamine: A connectionist approach to behavior and biology in schizophrenia. Psychological Review. 1992; 99(1):45–77. https://doi.org/10.1037/0033-295x.99.1.45 PMID: 1546118

20. Maniscalco B, Lau H. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. Consciousness and Cognition. 2012; 21(1):422–30. https://doi.org/10.1016/j.concog.2011.09.021 PMID: 22071269

21. Fleming SM. HMeta-d: hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. Neurosci Conscious. 2017; 2017(1):nix007–nix. https://doi.org/10.1093/nc/nix007 PMID: 29877507.

22. Howell D. Statistical methods for psychology: Cengage Learning. Inc, Belmont, CA. 2012:245.

23. Rounis E, Maniscalco B, Rothwell JC, Passingham RE, Lau H. Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. Cognitive Neuroscience. 2010; 1 (3):165–75. https://doi.org/10.1080/17588921003632529 PMID: 24168333

24. Fleming SM, Lau HC. How to measure metacognition. Front Hum Neurosci. 2014; 8:443–. https://doi.org/10.3389/fnhum.2014.00443 PMID: 25076880.

25. Del Cul A, Dehaene S, Reyes P, Bravo E, Slachevsky A. Causal role of prefrontal cortex in the threshold for access to consciousness. Brain. 2009; 132(9):2531–40. https://doi.org/10.1093/brain/awp111 PMID: 19433438

26. Fleming SM, Weil RS, Nagy Z, Dolan RJ, Rees G. Relating introspective accuracy to individual differences in brain structure. Science. 2010; 329(5998):1541–3. https://doi.org/10.1126/science.1191883 PMID: 20847276.

27. Fleming SM, Dolan RJ. The neural basis of metacognitive ability. Philos Trans R Soc Lond B Biol Sci. 2012; 367(1594):1338–49. https://doi.org/10.1098/rstb.2011.0417 PMID: 22492751.

28. Fleming SM, Ryu J, Golfinos JG, Blackmon KE. Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. Brain. 2014; 137(Pt 10):2811–22. Epub 2014/08/06. https://doi.org/10.1093/brain/awu221 PMID: 25100039.

29. Lak A, Costa GM, Romberg E, Koulakov AA, Mainen ZF, Kepecs A. Orbitofrontal cortex is required for optimal waiting based on decision confidence. Neuron. 2014; 84(1):190–201. Epub 2014/09/18. https://doi.org/10.1016/j.neuron.2014.08.039 PMID: 25242219.

30. Bang D, Fleming SM. Distinct encoding of decision confidence in human medial prefrontal cortex. Proc Natl Acad Sci U S A. 2018; 115(23):6082–7. Epub 2018/05/21. https://doi.org/10.1073/pnas.1800795115 PMID: 29784814.

31. Miyamoto K, Osada T, Setsuie R, Takeda M, Tamura K, Adachi Y, et al. Causal neural network of metamemory for retrospection in primates. Science. 2017; 355(6321):188–93. https://doi.org/10.1126/science.aal0162 PMID: 28082592

32. Rouault M, Seow T, Gillan CM, Fleming SM. Psychiatric Symptom Dimensions Are Associated With Dissociable Shifts in Metacognition but Not Task Performance. Biological psychiatry. 2018; 84(6):443–51. Epub 2018/01/11. https://doi.org/10.1016/j.biopsych.2017.12.017 PMID: 29458997.

33. Moses-Payne ME, Rollwage M, Fleming SM, Roiser JP. Postdecision Evidence Integration and Depressive Symptoms. Front Psychiatry. 2019; 10:639–. https://doi.org/10.3389/fpsyt.2019.00639 PMID: 31607959.

34. Hoven M, Lebreton M, Engelmann JB, Denys D, Luigjes J, van Holst RJ. Abnormalities of confidence in psychiatry: an overview and future perspectives. Transl Psychiatry. 2019; 9(1):268–. https://doi.org/10.1038/s41398-019-0602-7 PMID: 31636252.

35. Seow TXF, Gillan CM. Transdiagnostic Phenotyping Reveals a Host of Metacognitive Deficits Implicated in Compulsivity. Sci Rep. 2020; 10(1):2883–. https://doi.org/10.1038/s41598-020-59646-4 PMID: 32076008.

36. Wong K-F, Wang X-J. A recurrent network mechanism of time integration in perceptual decisions. J Neurosci. 2006; 26(4):1314–28. https://doi.org/10.1523/JNEUROSCI.3733-05.2006 PMID: 16436619.

37. Roxin A, Ledberg A. Neurobiological models of two-choice decision making can be reduced to a one-dimensional nonlinear diffusion equation. PLoS Comput Biol. 2008; 4(3):e1000046–e. https://doi.org/10.1371/journal.pcbi.1000046 PMID: 18369436.

38. Wang X-J. Probabilistic Decision Making by Slow Reverberation in Cortical Circuits. Neuron. 2002; 36 (5):955–68. https://doi.org/10.1016/s0896-6273(02)01092-9 PMID: 12467598

39. Atiya NAA, Rañó I, Prasad G, Wong-Lin K. A neural circuit model of decision uncertainty and change-of-mind. Nat Commun. 2019; 10(1):2287–. https://doi.org/10.1038/s41467-019-10316-8 PMID: 31123260.

40. Atiya NAA, Zgonnikov A, O'Hora D, Schoemann M, Scherbaum S, Wong-Lin K. Changes-of-mind in the absence of new post-decision evidence. PLoS Comput Biol. 2020; 16(2):e1007149–e. https://doi.org/10.1371/journal.pcbi.1007149 PMID: 32012147.

41. Kepecs A, Uchida N, Zariwala HA, Mainen ZF. Neural correlates, computation and behavioural impact of decision confidence. Nature. 2008; 455(7210):227–31. https://doi.org/10.1038/nature07200 PMID: 18690210

42. Yeung N, Botvinick MM, Cohen JD. The Neural Basis of Error Detection: Conflict Monitoring and the Error-Related Negativity. Psychological Review. 2004; 111(4):931–59. https://doi.org/10.1037/0033-295x.111.4.939 PMID: 15482068

43. Nelson TO, Narens L. Why investigate metacognition. Metacognition: Knowing about knowing. 1994; 13:1–25.

44. Maniscalco B, Odegaard B, Grimaldi P, Cho SH, Basso MA, Lau H, et al. Tuned inhibition in perceptual decision-making circuits can explain seemingly suboptimal confidence behavior. PLoS Comput Biol. 2021; 17(3):e1008779–e. https://doi.org/10.1371/journal.pcbi.1008779 PMID: 33780449.

45. Pleskac TJ, Busemeyer JR. Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. Psychological Review. 2010; 117(3):864–901. https://doi.org/10.1037/a0019737 PMID: 20658856

46. Sanders JI, Hangya B, Kepecs A. Signatures of a Statistical Computation in the Human Sense of Confidence. Neuron. 2016; 90(3):499–506. https://doi.org/10.1016/j.neuron.2016.03.025 PMID: 27151640.

47. Niyogi RK, Wong-Lin K. Dynamic excitatory and inhibitory gain modulation can produce flexible, robust and optimal decision-making. PLoS Comput Biol. 2013; 9(6):e1003099–e. Epub 2013/06/27. https://doi.org/10.1371/journal.pcbi.1003099 PMID: 23825935.

48. Smith PL, Ratcliff R, Wolfgang BJ. Attention orienting and the time course of perceptual decisions: response time distributions with masked and unmasked displays. Vision Research. 2004; 44(12):1297–320. https://doi.org/10.1016/j.visres.2004.01.002 PMID: 15066392

49. Ditterich J. Evidence for time-variant decision making. European Journal of Neuroscience. 2006; 24 (12):3628–41. https://doi.org/10.1111/j.1460-9568.2006.05221.x PMID: 17229111

50. Churchland AK, Kiani R, Shadlen MN. Decision-making with multiple alternatives. Nat Neurosci. 2008; 11(6):693–702. Epub 2008/05/18. https://doi.org/10.1038/nn.2123 PMID: 18488024.

51. Drugowitsch J, Moreno-Bote R, Churchland AK, Shadlen MN, Pouget A. The cost of accumulating evidence in perceptual decision making. J Neurosci. 2012; 32(11):3612–28. https://doi.org/10.1523/JNEUROSCI.4010-11.2012 PMID: 22423085.

52. Bogacz R, Cohen JD. Parameterization of connectionist models. Behavior Research Methods, Instruments, & Computers. 2004; 36(4):732–41. https://doi.org/10.3758/bf03206554 PMID: 15641419

53. Rowan TH. Functional stability analysis of numerical algorithms: The University of Texas at Austin; 1990.

54. Nelder JA, Mead R. A Simplex Method for Function Minimization. The Computer Journal. 1965; 7 (4):308–13. https://doi.org/10.1093/comjnl/7.4.308

55. Shadlen MN, Newsome WT. Neural Basis of a Perceptual Decision in the Parietal Cortex (Area LIP) of the Rhesus Monkey. Journal of Neurophysiology. 2001; 86(4):1916–36. https://doi.org/10.1152/jn.2001.86.4.1916 PMID: 11600651

56. Roitman JD, Shadlen MN. Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. J Neurosci. 2002; 22(21):9475–89. https://doi.org/10.1523/JNEUROSCI.22-21-09475.2002 PMID: 12417672.

57. Kiani R, Corthell L, Shadlen MN. Choice certainty is informed by both evidence and decision time. Neuron. 2014; 84(6):1329–42. https://doi.org/10.1016/j.neuron.2014.12.015 PMID: 25521381.

58. Zung WWK. A Self-Rating Depression Scale. Archives of General Psychiatry. 1965; 12(1):63. https://doi.org/10.1001/archpsyc.1965.01720310065008 PMID: 14221692

59. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A Brief Measure for Assessing Generalized Anxiety Disorder. Archives of Internal Medicine. 2006; 166(10):1092. https://doi.org/10.1001/archinte.166.10.1092 PMID: 16717171

60. Mason O, Linney Y, Claridge G. Short scales for measuring schizotypy. Schizophrenia Research. 2005; 78(2–3):293–6. https://doi.org/10.1016/j.schres.2005.06.020 PMID: 16054803

61. Patton JH, Stanford MS, Barratt ES. Factor structure of the barratt impulsiveness scale. Journal of Clinical Psychology. 1995; 51(6):768–74. https://doi.org/10.1002/1097-4679(199511)51:6<768::aid-jclp2270510607>3.0.co;2-1 PMID: 8778124

62. Foa EB, Huppert JD, Leiberg S, Langner R, Kichic R, Hajcak G, et al. The Obsessive-Compulsive Inventory: Development and validation of a short version. Psychological Assessment. 2002; 14(4):485–96. https://doi.org/10.1037/1040-3590.14.4.485 PMID: 12501574

63. Liebowitz MR. Social Phobia. Archives of General Psychiatry. 1985; 42(7):729. https://doi.org/10.1001/archpsyc.1985.01790300097013 PMID: 2861796

64. Spielberger CD. State-Trait Anxiety Inventory for Adults. PsycTESTS Dataset: American Psychological Association (APA); 1983.

65. Saunders JB, Aasland OG, Babor TF, De La Fuente JR, Grant M. Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO Collaborative Project on Early Detection of Persons with Harmful Alcohol Consumption-II. Addiction. 1993; 88(6):791–804. https://doi.org/10.1111/j.1360-0443.1993.tb02093.x PMID: 8329970

66. Marin RS, Biedrzycki RC, Firinciogullari S. Reliability and validity of the apathy evaluation scale. Psychiatry Research. 1991; 38(2):143–62. https://doi.org/10.1016/0165-1781(91)90040-v PMID: 1754629

67. Garner DM, Olmsted MP, Bohr Y, Garfinkel PE. The Eating Attitudes Test: psychometric features and clinical correlates. Psychological Medicine. 1982; 12(4):871–8. https://doi.org/10.1017/s0033291700049163 PMID: 6961471

68. Harrison OK, Garfinkel SN, Marlow L, Finnegan S, Marino S, Nanz L, et al. The Filter Detection Task for measurement of breathing-related interoception and metacognition. BioRxiv. 2020.

69. Rouault M, Fleming SM. Formation of global self-beliefs in the human brain. Proceedings of the National Academy of Sciences. 2020; 117(44):27268–76. https://doi.org/10.1073/pnas.2003094117 PMID: 33060292

70. Vaghi MM, Luyckx F, Sule A, Fineberg NA, Robbins TW, De Martino B. Compulsivity Reveals a Novel Dissociation between Action and Confidence. Neuron. 2017; 96(2):348–54.e4. Epub 2017/09/28. https://doi.org/10.1016/j.neuron.2017.09.006 PMID: 28965997.

71. Koren D, Poyurovsky M, Seidman LJ, Goldsmith M, Wenger S, Klein EM. The neuropsychological basis of competence to consent in first-episode schizophrenia: A pilot metacognitive study. Biological Psychiatry. 2005; 57(6):609–16. https://doi.org/10.1016/j.biopsych.2004.11.029 PMID: 15780847

72. Hauser TU, Allen M, Consortium N, Rees G, Dolan RJ. Metacognitive impairments extend perceptual decision making weaknesses in compulsivity. Sci Rep. 2017; 7(1):6614–. https://doi.org/10.1038/s41598-017-06116-z PMID: 28747627.

73. Hauser TU, Moutoussis M, Consortium N, Dayan P, Dolan RJ. Increased decision thresholds trigger extended information gathering across the compulsivity spectrum. Transl Psychiatry. 2017; 7(12):1296–. https://doi.org/10.1038/s41398-017-0040-3 PMID: 29249811.

74. Brown HR, Zeidman P, Smittenaar P, Adams RA, McNab F, Rutledge RB, et al. Crowdsourcing for cognitive science—the utility of smartphones. PLoS One. 2014; 9(7):e100662–e. https://doi.org/10.1371/journal.pone.0100662 PMID: 25025865.

75. Stolyarova A, Rakhshan M, Hart EE, O'Dell TJ, Peters MAK, Lau H, et al. Contributions of anterior cingulate cortex and basolateral amygdala to decision confidence and learning under uncertainty. Nat Commun. 2019; 10(1):4704–. https://doi.org/10.1038/s41467-019-12725-1 PMID: 31624264.

76. Adler WT, Ma WJ. Limitations of Proposed Signatures of Bayesian Confidence. Neural Computation. 2018; 30(12):3327–54. https://doi.org/10.1162/neco_a_01141 PMID: 30314423

77. Rausch M, Hellmann S, Zehetleitner M. Confidence in masked orientation judgments is informed by both evidence and visibility. Attention, Perception, & Psychophysics. 2017; 80(1):134–54. https://doi.org/10.3758/s13414-017-1431-5 PMID: 29043657

78. Resulaj A, Kiani R, Wolpert DM, Shadlen MN. Changes of mind in decision-making. Nature. 2009; 461(7261):263–6. https://doi.org/10.1038/nature08275 PMID: 19693010

79. Albantakis L, Deco G. Changes of mind in an attractor network of decision-making. PLoS Comput Biol. 2011; 7(6):e1002086. https://doi.org/10.1371/journal.pcbi.1002086 PMID: 21731482

80. Berlemont K, Martin J-R, Sackur J, Nadal J-P. Nonlinear neural network dynamics accounts for human confidence in a sequence of perceptual decisions. Sci Rep. 2020; 10(1):7940–. https://doi.org/10.1038/s41598-020-63582-8 PMID: 32409634.

81. García-Pérez MA. Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties. Vision Research. 1998; 38(12):1861–81. https://doi.org/10.1016/s0042-6989(97)00340-4 PMID: 9797963

82. de Leeuw JR. jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. Behavior Research Methods. 2014; 47(1):1–12. https://doi.org/10.3758/s13428-014-0458-y PMID: 24683129

83. Gillan CM, Kosinski M, Whelan R, Phelps EA, Daw ND. Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. Elife. 2016; 5:e11305. https://doi.org/10.7554/eLife.11305 PMID: 26928075.