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

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

1 Max Planck University College London Centre for Computational Psychiatry and Ageing Research, University College London, London WC1B 5EH, UK
2 Wellcome Centre for Human Neuroimaging, University College London, 12 Queen Square, London WC1N 3BG, UK
3 Department of Experimental Psychology, University College London, 26 Bedford Way, London WC1H 0AP, UK
4 Division of Psychiatry, University College London, 149 Tottenham Court Road, London W1T 7NF, UK

Corresponding Author: Nadim A. A. Atiya (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. In contrast to existing theoretical work, we account for empirical confidence judgements by fitting our biophysical model solely 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.

Introduction

Computational psychiatry (Friston et al. 2014; Huys, Maia, and Frank 2016; Wang and Krystal 2014; Montague et al. 2012) employs mechanistic and theory-driven models to relate brain function to phenomena that characterise mental health disorders (Ratcliff 1978; Ratcliff, Smith, and McKoon 2015; Rescorla, Wagner, and others 1972; Huys, Maia, and Frank 2016; Sutton and Barto 2018). Typically, algorithmic-level models (Marr and Poggio 1976) 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; Sutton and Barto 2018) closely relate to the functions of discrete brain circuits (Schultz 1999; Dayan and Balleine 2002; Dolan and Dayan 2012). 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 (Stephan et al. 2016). Here detailed biophysical models (Murray et al. 2014; Krystal et al. 2017; Rolls, Loh, and Deco 2008) may provide tools for understanding mental health disorders in terms of precise disturbances at the microcircuit level. For instance, Murray et al. (2014) 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 (Dima et al. 2009; Yang et al. 2014) have been employed to relate macrocircuit-level dysfunctions to symptoms of mental health disorders, and motivate non-invasive experimental neuroimaging to probe such dysfunctions (Cohen and Servan-Schreiber 1992). 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 (Maniscalco and Lau 2012; Fleming 2017). 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 (Howell 2009; Rounis et al. 2010). The ratio $meta - d'/d'$ thus yields a measure of metacognitive efficiency, i.e. metacognitive sensitivity for a given level of task performance (Fleming and Lau 2014).

Experimental evidence suggests that these facets of metacognitive ability are dissociable from task performance, and may have a distinct neural and computational basis (Del Cul et al. 2009; Fleming et al. 2010; Fleming and Dolan 2012). Interestingly, self-reported mental health symptoms have been linked to changes in metacognition, often in the absence of differences in task performance (Rouault et al. 2018; Moses-Payne et al. 2019; Hoven et al., 2019; Seow & Gillan, 2020). 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 (Wong and Wang 2006; Roxin and Ledberg 2008) the provide a detailed biophysical model of decision making (Wang 2002), enabling more rigorous theoretical analyses and faster computation. For instance, Wong and Wang (2006) have accounted for most of the behavioural results addressed by Wang (2002)'s model using the two slowest N-Methyl-D-aspartic acid (NMDA) dynamical variables. More recently, Atiya et al. (2019) extended Wong and Wang (2006)'s model to account for decision confidence reports and other metacognitive behaviours, such as an ability to flexibly change one's mind and correct errors (Atiya et al. 2020). More specifically, guided by neurophysiological evidence that supports an encoding of confidence within higher-order prefrontal brain regions (Kepecs et al. 2008; Fleming and Dolan 2012), 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 (Yeung et al., 2004).

To gain insight into potential mechanisms underlying shifts in metacognition, we first demonstrate that our biologically-motivated model (Atiya et al. 2019, 2020) can account for human confidence reports. Crucially, we show that the intrinsic dynamics of the model, constrained by first-order performance, are sufficient to account for subjects' confidence reports, going beyond existing methods of fitting models to empirical confidence data (Kepecs et al. 2008; Kiani and Shadlen 2009; Pleskac and Busemeyer 2010; Sanders, Hangya, and Kepecs 2016). 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 uncertainty monitoring. Finally, following a computational psychiatry approach we show that disturbances in uncertainty monitoring can be associated with variation in self-reported psychopathology. Our work provides a computational framework for mapping theoretical measurements 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 (Wong and Wang 2006). Importantly, our model builds on neurophysiological evidence suggesting that decision confidence is encoded by dedicated higher-order brain regions (Kepecs et al. 2008; Fleming and Dolan 2012). 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 by integrating the summed neuronal activities of sensorimotor populations (see Fig. 1C for a sample trial). Importantly, 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 (Niyogi and Wong-Lin 2013; Smith, Ratcliff, and Wolfgang 2004; Ditterich 2006; Churchland, Kiani, and Shadlen 2008; Kiani, Hanks, and Shadlen 2008; Drugowitsch et al. 2012). Here we refer to this feedback loop as the strength of uncertainty-monitoring (UM).

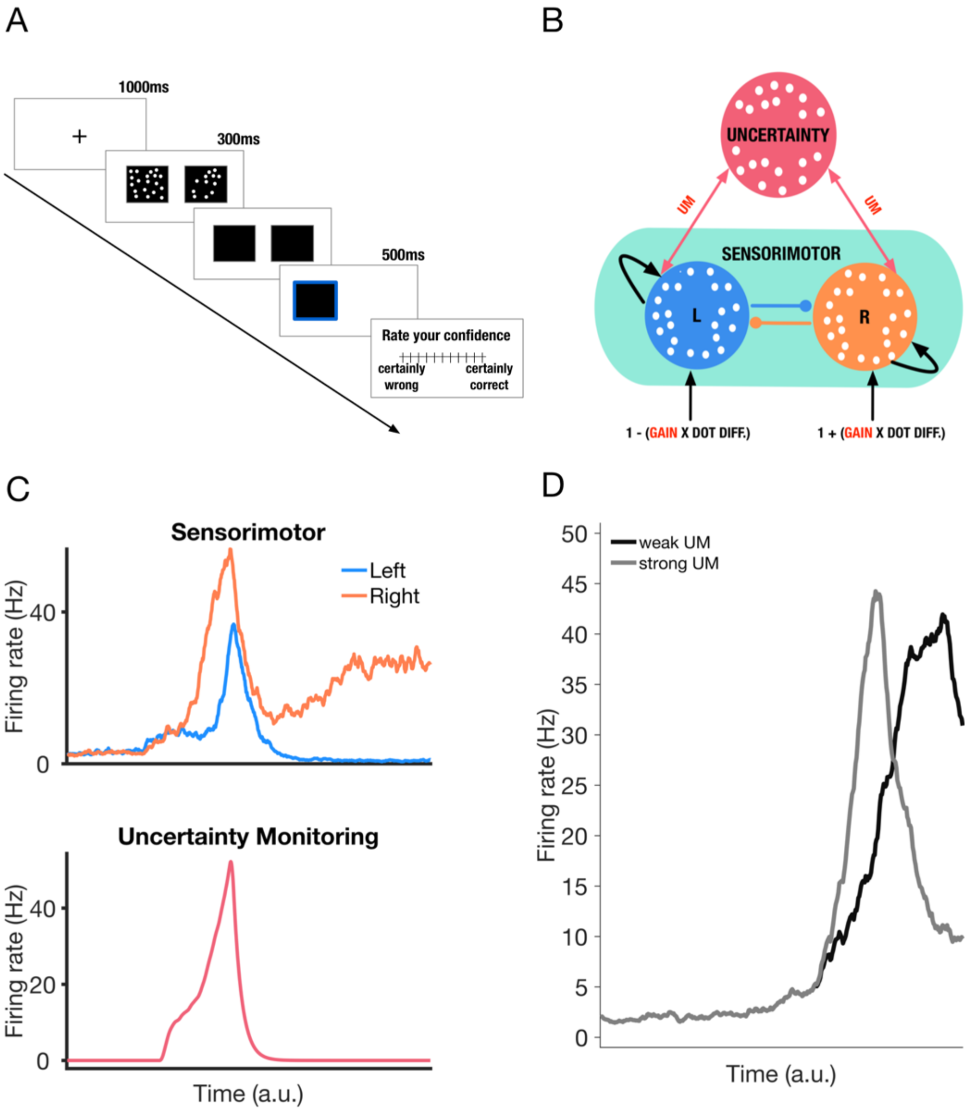

Figure 1. Task and neural circuit model. A. Perceptual decision-making task used as a basis for simulations. A fixation cross appears for 1000ms, followed by two boxes with dots for a fixed duration of 300ms. Subjects are asked to judge which

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 (Supplementary Notes 1 and 2). **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 monitoring 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 Atiya et al. (2019, 2020) 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-monitoring (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).

Applying the model to account for facets of metacognition

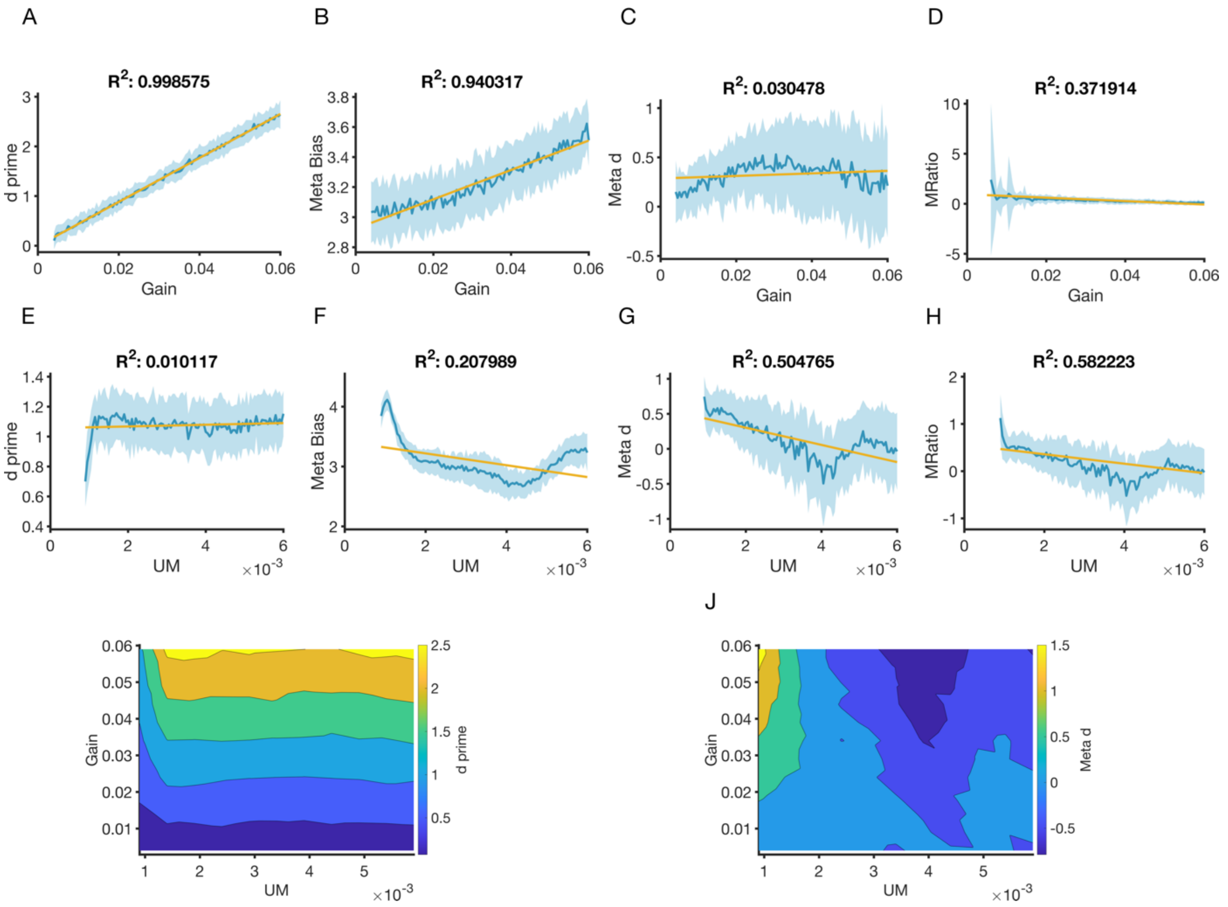

Figure 2. Dissociable changes in metacognition are associated with changes in uncertainty monitoring. The behaviour of the model was analysed using standard metrics of performance (d') and metacognition (metacognitive bias, sensitivity (meta-d') and efficiency (meta-d'/d')). Blue line represents mean 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' and (**B**) metacognitive bias but (**C**) almost no effect on sensitivity. Gain has a weak negative effect on (**D**) metacognitive efficiency, possibly driven by the strong linear increase in d' in panel A. Increasing UM has no effect on (**E**) d', but a negative effect on (**F**) metacognitive bias, (**G**) metacognitive sensitivity, and (**H**) metacognitive efficiency. 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 0.0015 (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).

We first asked whether our model can account for the 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 (Atiya et al., 2019). 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 metacognition (Fleming 2017). 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, Maniscalco and Lau 2012). 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 (Figs. 2A and 2B) show increasing gain has a strong positive effect on d' and metacognitive bias (i.e. leading to higher confidence levels). The effect on d' is unsurprising given that increasing gain magnifies the difference in input each neuronal population is receiving (see Fig. 1B). The effect on metacognitive bias is also expected from an overall increase in number of correct trials, and therefore the production of fewer 'low-confidence' error trials. Notably, however, the results also show (Fig. 2C) that increasing gain has almost no effect on $meta - d'$. Finally, the results (Fig. 2D) show that increasing gain has a weak negative 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 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 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.

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 monitoring 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

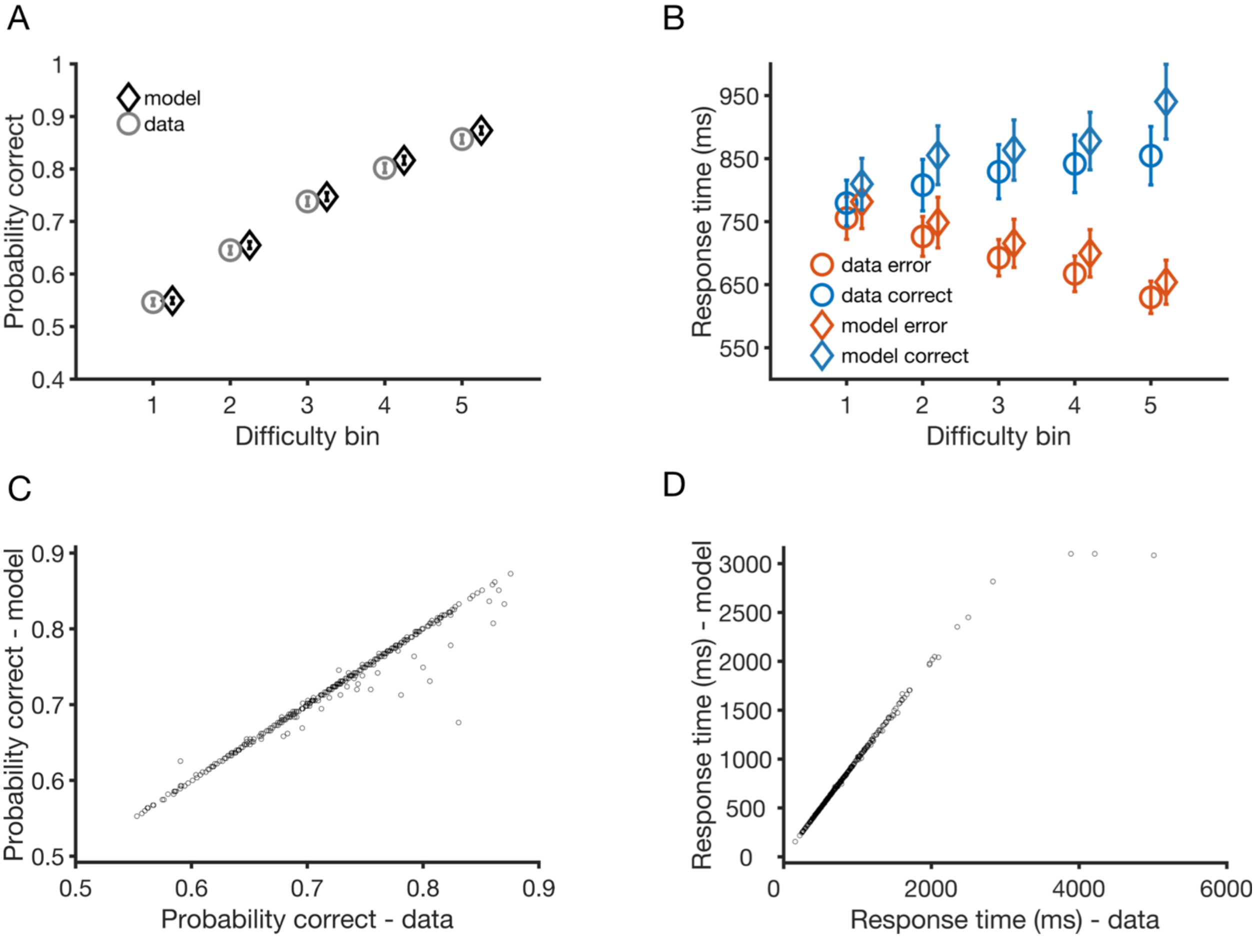

Figure 3. Model accounts for subjects' perceptual performance in experiment 1. A. Choice accuracy, i.e. probability correct as a function of task difficulty from experiment 1 of Rouault et al. (2018) averaged across all 498 participants. Task difficulty is split into 5 difficulty bins as in the original 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 mean overall response times and **D.** overall accuracy 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 Figures **C** and **D** (but not **A** and **B**). See Supplementary Figure 8 for scatter plots without resetting the random generator seed.

We re-analysed data from Rouault et al. (2018), 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 (Atiya et al. 2019, 2020). 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 Atiya et al. (2019, 2020) (See Fig. 1B).

In fitting our model to subjects' choices and response times, we used a procedure based on the subplex optimisation method (Bogacz and Cohen 2004; Rowan 1990) (see Methods). The subplex optimisation method is an evolution of the simplex method (Nelder and Mead 1965) – 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 (Atiya et al. 2020), 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 of the gain parameter generally have better choice accuracy. The second parameter is the strength of uncertainty monitoring (see Fig. 2D 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 (Shadlen and Newsome 2001; Roitman and Shadlen 2002; Sanders, Hangya, and Kepecs 2016), subjects' correct (error) responses were quicker (slower) as the task became easier, forming a '<' pattern of response times as a function of difficulty (see Fig. 3B, and Fig. 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 (Kepecs et al. 2008; Kiani, Corthell, and Shadlen 2014). In our model, longer response times allow more time for the uncertainty monitoring population to activate — leading to higher uncertainty (see Methods).

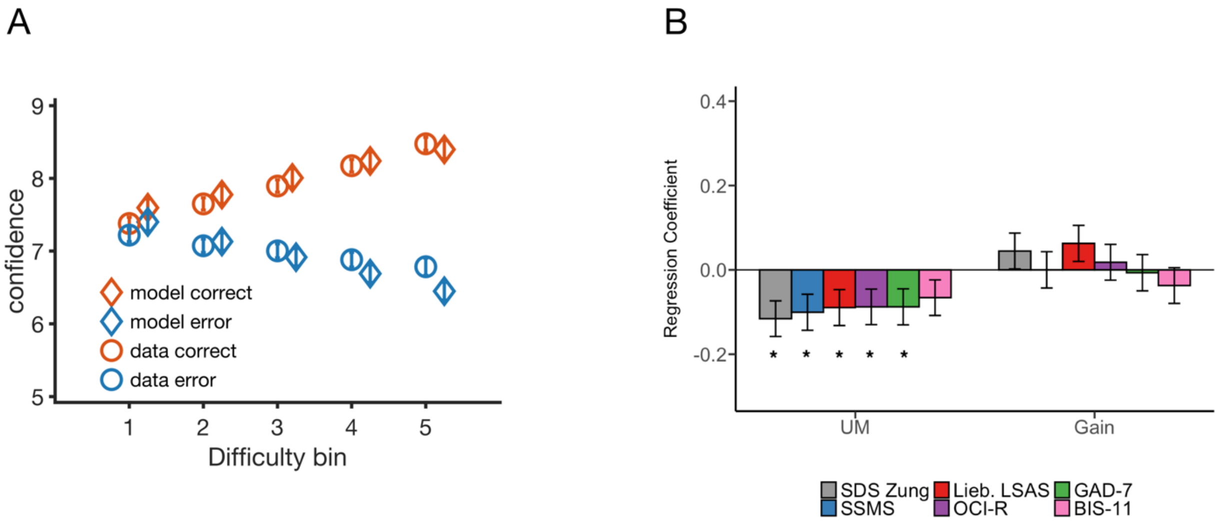

Figure 4. Model accounts for subjects' confidence reports and individual differences in uncertainty monitoring predict symptom scores. A. Confidence reports averaged across all participants from experiment 1 data (circles) and model (diamond) as a function of task difficulty. Blue (orange) markers: Error (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-monitoring 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-monitoring. No significant association was found between impulsivity (pink) and the strength of uncertainty-monitoring. 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 Supplementary Figure 6 for regression model results with age and IQ predicting model parameters). * p< 0.05.

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 (Sanders, Hangya, and Kepecs 2016) 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 a 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. Hence the model is able to account for individual differences in subjects' perceptual and metacognitive performance despite model fits only having access to choices and response times. We next asked whether the uncertainty-monitoring mechanism 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 (Zung 1965; Spitzer et al. 2006; Mason, Linney, and Claridge 2005; Patton, Stanford, and Barratt 1995; Foa et al. 2002; Liebowitz et al. 1985; Spielberger and Gorsuch 1983; Saunders et al. 1993; Marin, Biedrzycki, and Firinciogullari 1991; Garner et al. 1982). 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 Rouault et al. (2018), 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 monitoring. 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.

We entered each questionnaire score (see Methods) into multiple linear regressions predicting the uncertainty-monitoring and gain parameters. The results (see Fig. 4B) show that increases in z-scored self-reported scores were broadly associated with weaker uncertainty monitoring 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 Rouault et al. (2018), but now provide a potential circuit-level explanation for such differences (i.e., a change in the strength of uncertainty-monitoring).

We also followed the same approach for experiment 2 (see Supplementary Figure Note 2), although here we found no significant association between the majority of self-reported scores (or cross-cutting factors derived from these scores, see Supplementary Figures 3A and 3B) 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-monitoring being less constrained by the data (see Supplementary Figure 7). To explore this further, we attempted to recover the fitted parameters to both experiment 1 and 2 data and found that the results show that the fits to experiment 1 data were indeed more stable (See Supplementary Figures 4 and 5) – 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 monitoring in experiment 1 were also negatively related to the uncertainty monitoring in experiment 2.

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, stronger uncertainty monitoring is associated with decreased metacognitive bias, sensitivity, and efficiency. Importantly, changes in uncertainty-monitoring strength have no effect on perceptual sensitivity. Notably, our model-fitting approach enabled inferences about uncertainty monitoring (and, in turn, these facets of metacognition) from fits to first-order performance data alone. 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 monitoring.

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 (Kepecs et al. 2008; Kiani, Corthell, and Shadlen 2014; Atiya et al. 2020). 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 (Atiya et al. 2019). By linking our model's uncertainty to standard metrics of metacognition, we reveal that shifts in facets of metacognition are associated with disturbances in uncertainty monitoring (Fig. 2). This suggest that stronger uncertainty monitoring in such a network has a negative effect on metacognitive bias, sensitivity, efficiency, while leaving perceptual sensitivity unaffected. More specifically, controlling for task difficulty, our findings reveal that stronger uncertainty-monitoring (i.e. leading to overall faster responses) leads to deficits in the accuracy of confidence reports – generally leading to lower confidence in correct trials, and higher confidence in errors. 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-monitoring, finds support in recent neuropsychological work. For instance, lesions in prefrontal brain regions are associated with deficits in metacognitive ability, but not task performance (Fleming et al., 2014), highlighting the contribution of higher-order brain regions to metacognition (Fleming et al. 2010; Fleming and Dolan 2012).

Future work could combine our computational framework with neuroimaging to further elucidate the neural basis of metacognitive ability.

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 (Rouault et al. 2018). Rouault and colleagues 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 monitoring, in the absence of any change in sensory gain. Our analyses not only recapitulate previously-reported relationships with depression and anxiety (Fig. 5B), but show that schizotopy and OCD scores also relate to disturbances in uncertainty-monitoring (Vaghi et al., 2017), in line with existing work relating deficits in self-evaluation to schizophrenia (Koren et al. 2004).

Symptoms of OCD have been linked to deficits in working memory (Nakao et al. 2009), though previous work has linked the typical feeling of doubt in OCD patients to an intolerance of uncertainty (Tolin et al. 2003). More recent work has demonstrated that symptoms of OCD are associated with deficits in utilising evidence to update confidence (Seow & Gillan, 2020). 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 high 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 (Hauser, Moutoussis, et al. 2017; Hauser, Allen, et al. 2017). 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 Supplementary Note 1 and 2 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 (Brown et al. 2014).

We also explored whether our model accounts for metacognition-psychopathology relationships in a task with staircased difficulty levels (experiment 2 in Rouault et al.). Although our analyses of the UM parameter show a similar pattern to those obtained for metacognitive bias in the original study, (Supplementary Figures 3A and 3B), 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-monitoring 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 (Supplementary Figures 4 and 5). Further theoretical work is needed to determine the effect of per-subject difficulty variance on the ability to infer such model parameters.

Previous versions of our neural circuit model have also been applied to tasks with explicit motor reaching trajectories through a dedicated motor output network (Atiya et al. 2019, 2020). 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 monitoring. 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

Neural circuit model of uncertainty

We modelled the processes underpinning decisions and confidence using a neural circuit model of uncertainty described previously (Atiya et al. 2019, 2020). 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 (Atiya et al. 2019, 2020), the sensorimotor module is modelled using a reduced (i.e. two-variable) spiking neural network model (Wang 2002; Wong and Wang 2006). The dynamics of the neuronal populations are described by:

$$\frac{\mathrm{d}S_L}{\mathrm{d}t} = -\frac{S_L}{\tau_s} + (1 - S_L)\gamma H(x_L, x_R) \qquad \textit{Eq. 1}$$

$$\frac{\mathrm{d}S_R}{\mathrm{d}t} = -\frac{S_R}{\tau_s} + (1 - S_R)\gamma H(x_R, x_L) \qquad \textit{Eq. 2}$$

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 (Wong and Wang 2006) that describes a reduction of the original spiking neuronal network model of decision making (Wang 2002).

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

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

where a, b, d are parameters fitted to the leaky integrate-and-fire model (Wang 2002). 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:

$$x_i = w_+ S_i - w_- S_j + I_c + I_i + I_\sigma + w_u U \qquad Eq.\ 4$$

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 (Wong and Wang 2006). 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:

$$I_i = w_e \mu_0\ (1 \pm \varepsilon) \qquad Eq.\ 5$$

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

$$\varepsilon = \text{gain} \cdot \text{dot difference} \qquad Eq.\ 6$$

where the input gain parameter maps the dot difference to difference in input flowing into the sensorimotor populations.

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-monitoring strength. U denotes the dynamical variable of the uncertainty-monitoring population, which is described by:

$$\tau_u \frac{dU}{dt} = [H_L + H_R - l]_+ - U \qquad \textit{Eq. 7}$$

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 2). We summarise the values of all model parameters in Table 1.

Quantifying uncertainty within a trial

As in our previous work (Atiya et al. 2020), 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; Sanders, Hangya, and Kepecs 2016). 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 (Bogacz and Cohen 2004; Rowan 1990). Subplex optimisation is based on the simplex optimsation method, but adapted for noisy objective functions (Rowan 1990). For each participant, we minimise the cost function:

$$\text{cost} \;=\; \frac{1}{m}(\text{RT}_{\text{model}} - \text{RT}_{\text{data}})^2 + \frac{1}{n}(\text{accuracy}_{\text{model}} - \text{accuracy}_{\text{data}})^2 \qquad \textit{Eq. 8}$$

where $\mathrm{RT}_{\mathrm{model}}$ is the model's mean response time from a single model simulation (with a fixed random seed), $\mathrm{RT}_{\mathrm{data}}$denotes the participants' mean response time. Similarly, $\mathrm{accuracy}_{\mathrm{model}}$ and $\mathrm{accuracy}_{\mathrm{data}}$ 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., $m = \mathrm{RT}_{\mathrm{model}}$, and $n = \mathrm{accuracy}_{\mathrm{model}}$) (Bogacz & Cohen, 2002). Importantly, we only fit two free parameters: gain and $\mathrm{w_u}$, from Eqs. 6 and 4, respectively. The vast majority of the other model parameters are adapted from our previous work (Atiya et al. 2020) (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.

Ethics statement

Data analysed in this work was first collected as part of a study conducted by Rouault et al. (2018). Participants provided consent in accordance with procedures approved by the University College London Research Ethics Committee (Project ID 1260/003).

Participants

We re-analysed data from Rouault et al. (2018), 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 pas paid if participants achieved task performance between 60-85%, and passed a check question. We used the same exclusion criteria applied in Rouault et al. (2018) and described in the Supplementary Material of that paper.

Task

In both experiments, participants completed a simple perceptual decision-making task where they judged which box contained a higher number of dots, with no feedback. 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. After every trial particpants provided a confidence judgement on a full 11-point probabilistic scale (Boldt and Yeung 2015): 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.

Experiment 2 (see Supplementary Note 1) is identical to experiment 1 in all but three aspects. First, Rouault et al. (2018) used a staircase (calibration) procedure to fix participants' perceptual performance (Garcıa-Pérez 1998; Fleming et al. 2010). 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.

Psychiatric questionnaires

Participants completed a set of self-report questionnaires used to assess their psychiatric symptoms (Rouault et al. 2018). In experiment 1, the questionnaires were:

- Depression using the Self-Rating Depression Scale (SDS) (Zung 1965).
- Generalised anxiety using the Generalised Anxiety Disorder 7-Item Scale (GAD-7) (Spitzer et al. 2006)
- Schizotypy using the Short Scales for Measuring Schizotypy (SSMS) (Mason, Linney, and Claridge 2005)
- Impulsivity using the Barratt Impulsiveness Scale (BIS-11) (Patton, Stanford, and Barratt 1995)
- Obsessive Compulsive Disorder (OCD) using the Obsessive-Compulsive Inventory–Revised (OCI-R) (Foa et al. 2002)
- Social anxiety using the Liebowitz Social Anxiety Scale (LSAS) (Liebowitz et al. 1985)

In experiment 2 (Supplementary Notes 1 and 2), 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 (Spielberger and Gorsuch 1983)
- Alcoholism was assessed with the Alcohol Use Disorders Identification Test (AUDIT) (Saunders et al. 1993)
- Apathy was assessed with the Apathy Evaluation Scale (AES) (Marin, Biedrzycki, and Firinciogullari 1991)

- Eating disorders was assessed with the Eating Attitudes Test (EAT-26) (Garner et al. 1982)

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 (Gillan et al. 2016).

Factor analysis

For experiment 2 data (see Supplementary Notes 1 and 2), we obtained three latent factors that explain the shared variance across the 209 questionnaire items. To do that, we followed the same approach in Rouault et al. (2018) and Gillan et al. (2016), 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 Rouault et al. (2018). 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:

$$\mathrm{Param} = \beta_0 + \beta_1\mathrm{Score} + \beta_2\mathrm{Age} + \beta_3\mathrm{Gender} + \beta_4\mathrm{IQ} \qquad \textit{Eq. 9}$$

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

$$\mathrm{Param} = \beta_0 + \beta_1\mathrm{Factor\ 1} + \beta_2\mathrm{Factor\ 2} + \beta_3\mathrm{Factor\ 3} + \beta_4\mathrm{Age} + \beta_5\mathrm{Gender} + \beta_6\mathrm{IQ} \qquad \textit{Eq.10}$$

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:

$$\mathrm{metric} = \beta_0 + \beta_1\mathrm{model\ param} \qquad \textit{Eq.11}$$

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 (Fleming 2017). The model returns a value of metacognitive sensitivity ($meta - d'$) for each simulated dataset. To compute metacognitive efficiency, we calculated the ratio $meta - d'/d'$.

Table 1. Table of fixed model parameter values for all participants. Parameters $\tau_S, \tau_u, a, b, d, I_c, w_e$ were directly adapted from (Atiya et al. 2020). 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}

Data availability

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

Acknowledgments

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.

References

Atiya, Nadim AA, Iñaki Rañó, Girijesh Prasad, and KongFatt Wong-Lin. 2019. "A Neural Circuit Model of Decision Uncertainty and Change-of-Mind." *Nature Communications* 10 (1): 1–12.

Atiya, Nadim AA, Arkady Zgonnikov, Denis O'Hora, Martin Schoemann, Stefan Scherbaum, and KongFatt Wong-Lin. 2020. "Changes-of-Mind in the Absence of New Post-Decision Evidence." *PLOS Computational Biology* 16 (2): e1007149.

Bogacz, Rafal, and Jonathan D Cohen. 2004. "Parameterization of Connectionist Models." *Behavior Research Methods, Instruments, & Computers* 36 (4): 732–41.

Boldt, Annika, and Nick Yeung. 2015. "Shared Neural Markers of Decision Confidence and Error Detection." *Journal of Neuroscience* 35 (8): 3478–84.

Brown, Harriet R, Peter Zeidman, Peter Smittenaar, Rick A Adams, Fiona McNab, Robb B Rutledge, and Raymond J Dolan. 2014. "Crowdsourcing for Cognitive Science–the Utility of Smartphones." *PloS One* 9 (7).

Churchland, Anne K, Roozbeh Kiani, and Michael N Shadlen. 2008. "Decision-Making with Multiple Alternatives." *Nature Neuroscience* 11 (6): 693.

Cohen, Jonathan D, and David Servan-Schreiber. 1992. "Context, Cortex, and Dopamine: A Connectionist Approach to Behavior and Biology in Schizophrenia." *Psychological Review* 99 (1): 45.

Condon, David M, and William Revelle. 2014. "The International Cognitive Ability Resource: Development and Initial Validation of a Public-Domain Measure." *Intelligence* 43: 52–64.

Dayan, Peter, and Bernard W Balleine. 2002. "Reward, Motivation, and Reinforcement Learning." *Neuron* 36 (2): 285–98.

Del Cul, Antoine, Stanislas Dehaene, P Reyes, E Bravo, and Andrea Slachevsky. 2009. "Causal role of prefrontal cortex in the threshold for access to consciousness." *Brain* 132 (9): 2531–40.

Dima, Danai, Jonathan P Roiser, Detlef E Dietrich, Catharina Bonnemann, Heinrich Lanfermann, Hinderk M Emrich, and Wolfgang Dillo. 2009. "Understanding Why Patients with Schizophrenia Do Not Perceive the Hollow-Mask Illusion Using Dynamic Causal Modelling." *Neuroimage* 46 (4): 1180–6.

Ditterich, Jochen. 2006. "Evidence for Time-Variant Decision Making." *European Journal of Neuroscience* 24 (12): 3628–41.

Dolan, Ray J, Peter Dayan (2013). Goals and habits in the brain. Neuron, 80(2), 312-325.

Drugowitsch, Jan, Rubén Moreno-Bote, Anne K Churchland, Michael N Shadlen, and Alexandre Pouget. 2012. "The Cost of Accumulating Evidence in Perceptual Decision Making." *Journal of Neuroscience* 32 (11): 3612–28.

Fleming, Stephen M. 2017. "HMeta-d: Hierarchical Bayesian Estimation of Metacognitive Efficiency from Confidence Ratings." *Neuroscience of Consciousness* 2017 (1): nix007.

Fleming, Stephen M, and Raymond J Dolan. 2012. "The Neural Basis of Metacognitive Ability." *Philosophical Transactions of the Royal Society B: Biological Sciences* 367 (1594): 1338–49.

Fleming, Stephen M, and Hakwan C Lau. 2014. "How to Measure Metacognition." *Frontiers in Human Neuroscience* 8: 443.

Fleming, S. M., Ryu, J., Golfinos, J. G., & Blackmon, K. E. (2014). Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. Brain : a journal of neurology, 137(Pt 10), 2811–2822.

Fleming, Stephen M, Rimona S Weil, Zoltan Nagy, Raymond J Dolan, and Geraint Rees. 2010. "Relating Introspective Accuracy to Individual Differences in Brain Structure." *Science* 329 (5998): 1541–3.

Foa, Edna B, Jonathan D Huppert, Susanne Leiberg, Robert Langner, Rafael Kichic, Greg Hajcak, and Paul M Salkovskis. 2002. "The Obsessive-Compulsive Inventory: Development and Validation of a Short Version." *Psychological Assessment* 14 (4): 485.

Friston, Karl J, Klaas Enno Stephan, Read Montague, and Raymond J Dolan. 2014. "Computational Psychiatry: The Brain as a Phantastic Organ." *The Lancet Psychiatry* 1 (2): 148–58.

Garcıa-Pérez, Miguel A. 1998. "Forced-Choice Staircases with Fixed Step Sizes: Asymptotic and Small-Sample Properties." *Vision Research* 38 (12): 1861–81.

Garner, David M, Marion P Olmsted, Yvonne Bohr, and Paul E Garfinkel. 1982. "The Eating Attitudes Test: Psychometric Features and Clinical Correlates." *Psychological Medicine* 12 (4): 871–78.

Gillan, Claire M, Michal Kosinski, Robert Whelan, Elizabeth A Phelps, and Nathaniel D Daw. 2016. "Characterizing a Psychiatric Symptom Dimension Related to Deficits in Goal-Directed Control." *Elife* 5: e11305.

Hauser, Tobias U, Micah Allen, Geraint Rees, and Raymond J Dolan. 2017. "Metacognitive Impairments Extend Perceptual Decision Making Weaknesses in Compulsivity." *Scientific Reports* 7 (1): 1–10.

Hauser, Tobias U, Michael Moutoussis, Peter Dayan, and Raymond J Dolan. 2017. "Increased Decision Thresholds Trigger Extended Information Gathering Across the Compulsivity Spectrum." *Translational Psychiatry* 7 (12): 1–10.

Hoven, M., Lebreton, M., Engelmann, J. B., Denys, D., Luigjes, J., & van Holst, R. J. (2019). Abnormalities of confidence in psychiatry: an overview and future perspectives. Translational psychiatry, 9(1), 1-18.

Howell, David C. 2009. *Statistical Methods for Psychology*. Cengage Learning.

Huys, Quentin JM, Tiago V Maia, and Michael J Frank. 2016. "Computational Psychiatry as a Bridge from Neuroscience to Clinical Applications." *Nature Neuroscience* 19 (3): 404.

Kepecs, Adam, Naoshige Uchida, Hatim A Zariwala, and Zachary F Mainen. 2008. "Neural Correlates, Computation and Behavioural Impact of Decision Confidence." *Nature* 455 (7210): 227–31.

Kiani, Roozbeh, Leah Corthell, and Michael N Shadlen. 2014. "Choice Certainty Is Informed by Both Evidence and Decision Time." *Neuron* 84 (6): 1329–42.

Kiani, Roozbeh, Timothy D Hanks, and Michael N Shadlen. 2008. "Bounded Integration in Parietal Cortex Underlies Decisions Even When Viewing Duration Is Dictated by the Environment." *Journal of Neuroscience* 28 (12): 3017–29.

Kiani, Roozbeh, and Michael N Shadlen. 2009. "Representation of Confidence Associated with a Decision by Neurons in the Parietal Cortex." *Science* 324 (5928): 759–64.

Koren, Danny, Larry J Seidman, Michael Poyurovsky, Morris Goldsmith, Polina Viksman, Suzi Zichel, and Ehud Klein. 2004. "The Neuropsychological Basis of Insight in First-Episode Schizophrenia: A Pilot Metacognitive Study." *Schizophrenia Research* 70 (2-3): 195–202.

Krystal, John H, John D Murray, Adam M Chekroud, Philip R Corlett, Genevieve Yang, Xiao-Jing Wang, and Alan Anticevic. 2017. "Computational Psychiatry and the Challenge of Schizophrenia." Oxford University Press US.

Liebowitz, Michael R, Jack M Gorman, Abby J Fyer, and Donald F Klein. 1985. "Social Phobia: Review of a Neglected Anxiety Disorder." *Archives of General Psychiatry* 42 (7): 729–36.

Maniscalco, Brian, and Hakwan Lau. 2012. "A Signal Detection Theoretic Approach for Estimating Metacognitive Sensitivity from Confidence Ratings." *Consciousness and Cognition* 21 (1): 422–30.

Marin, Robert S, Ruth C Biedrzycki, and Sekip Firinciogullari. 1991. "Reliability and Validity of the Apathy Evaluation Scale." *Psychiatry Research* 38 (2): 143–62.

Marr, David, and Tomaso Poggio. 1976. "From Understanding Computation to Understanding Neural Circuitry."

Mason, Oliver, Yvonne Linney, and Gordon Claridge. 2005. "Short Scales for Measuring Schizotypy." *Schizophrenia Research* 78 (2-3): 293–96.

Montague, P Read, Raymond J Dolan, Karl J Friston, and Peter Dayan. 2012. "Computational Psychiatry." *Trends in Cognitive Sciences* 16 (1): 72–80.

Moses-Payne, Madeleine E, Max Rollwage, Stephen M Fleming, and Jonathan P Roiser. 2019. "Post-Decision Evidence Integration and Depressive Symptoms." *Frontiers in Psychiatry* 10: 639.

Murray, John D, Alan Anticevic, Mark Gancsos, Megan Ichinose, Philip R Corlett, John H Krystal, and Xiao-Jing Wang. 2014. "Linking Microcircuit Dysfunction to Cognitive Impairment: Effects of Disinhibition Associated with Schizophrenia in a Cortical Working Memory Model." *Cerebral Cortex* 24 (4): 859–72.

Nakao, Tomohiro, Akiko Nakagawa, Eriko Nakatani, Maiko Nabeyama, Hirokuni Sanematsu, Takashi Yoshiura, Osamu Togao, et al. 2009. "Working Memory Dysfunction in Obsessive–Compulsive Disorder: A Neuropsychological and Functional Mri Study." *Journal of Psychiatric Research* 43 (8): 784–91.

Nedler, JA, and R Mead. 1965. "A Simplex Method for Function Minimization, Compt." *J* 7: 308–13.

Niyogi, Ritwik K, and KongFatt Wong-Lin. 2013. "Dynamic Excitatory and Inhibitory Gain Modulation Can Produce Flexible, Robust and Optimal Decision-Making." *PLoS Computational Biology* 9 (6).

Patton, Jim H, Matthew S Stanford, and Ernest S Barratt. 1995. "Factor Structure of the Barratt Impulsiveness Scale." *Journal of Clinical Psychology* 51 (6): 768–74.

Pleskac, Timothy J, and Jerome R Busemeyer. 2010. "Two-Stage Dynamic Signal Detection: A Theory of Choice, Decision Time, and Confidence." *Psychological Review* 117 (3): 864.

Ratcliff, Roger. 1978. "A Theory of Memory Retrieval." *Psychological Review* 85 (2): 59.

Ratcliff, Roger, Philip L Smith, and Gail McKoon. 2015. "Modeling regularities in response time and accuracy data with the diffusion model." *Current Directions in Psychological Science* 24 (6): 458–70.

Rescorla, Robert A, Allan R Wagner, and others. 1972. "A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement." *Classical Conditioning II: Current Research and Theory* 2: 64–99.

Roitman, Jamie D, and Michael N Shadlen. 2002. "Response of Neurons in the Lateral Intraparietal Area During a Combined Visual Discrimination Reaction Time Task." *Journal of Neuroscience* 22 (21): 9475–89.

Rolls, Edmund T, Marco Loh, and Gustavo Deco. 2008. "An Attractor Hypothesis of Obsessive–Compulsive Disorder." *European Journal of Neuroscience* 28 (4): 782–93.

Rouault, Marion, Tricia Seow, Claire M Gillan, and Stephen M Fleming. 2018. "Psychiatric Symptom Dimensions Are Associated with Dissociable Shifts in Metacognition but Not Task Performance." *Biological Psychiatry* 84 (6): 443–51.

Rounis, Elisabeth, Brian Maniscalco, John C Rothwell, Richard E Passingham, and Hakwan Lau. 2010. "Theta-Burst Transcranial Magnetic Stimulation to the Prefrontal Cortex Impairs Metacognitive Visual Awareness." *Cognitive Neuroscience* 1 (3): 165–75.

Rowan, T. 1990. "The Subplex Method for Unconstrained Optimization." PhD thesis, PhD thesis, Ph. D. thesis, Department of Computer Sciences, Univ. of Texas.

Roxin, Alex, and Anders Ledberg. 2008. "Neurobiological Models of Two-Choice Decision Making Can Be Reduced to a One-Dimensional Nonlinear Diffusion Equation." *PLoS Computational Biology* 4 (3).

Sanders, Joshua I, Balázs Hangya, and Adam Kepecs. 2016. "Signatures of a Statistical Computation in the Human Sense of Confidence." *Neuron* 90 (3): 499–506.

Saunders, John B, Olaf G Aasland, Thomas F Babor, Juan R De la Fuente, and Marcus Grant. 1993. "Development of the Alcohol Use Disorders Identification Test (Audit): WHO Collaborative Project on Early Detection of Persons with Harmful Alcohol Consumption-Ii." *Addiction* 88 (6): 791–804.

Schultz, Wolfram. 1999. "The Reward Signal of Midbrain Dopamine Neurons." *Physiology* 14 (6): 249–55.

Seow, T. X., & Gillan, C. M. (2020). Transdiagnostic Phenotyping Reveals a Host of Metacognitive Deficits Implicated in Compulsivity. Scientific reports, 10(1), 1-11.

Shadlen, Michael, and William Newsome. 2001. "Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey." *Journal of Neurophysiology* 86 (4): 1916–36.

Smith, Philip L, Roger Ratcliff, and Bradley J Wolfgang. 2004. "Attention Orienting and the Time Course of Perceptual Decisions: Response Time Distributions with Masked and Unmasked Displays." *Vision Research* 44 (12): 1297–1320.

Spielberger, Charles Donald, and Richard L Gorsuch. 1983. *State-Trait Anxiety Inventory for Adults: Manual and Sample: Manual, Instrument and Scoring Guide*. Consulting Psychologists Press.

Spitzer, Robert L, Kurt Kroenke, Janet BW Williams, and Bernd Löwe. 2006. "A Brief Measure for Assessing Generalized Anxiety Disorder: The Gad-7." *Archives of Internal Medicine* 166 (10): 1092–7.

Stephan, Klaas E, Dominik R Bach, Paul C Fletcher, Jonathan Flint, Michael J Frank, Karl J Friston, Andreas Heinz, et al. 2016. "Charting the Landscape of Priority Problems in Psychiatry, Part 1: Classification and Diagnosis." *The Lancet Psychiatry* 3 (1): 77–83.

Sutton, Richard S, and Andrew G Barto. 2018. *Reinforcement Learning: An Introduction*. MIT press.

Tolin, David F, Jonathan S Abramowitz, Bartholomew D Brigidi, and Edna B Foa. 2003. "Intolerance of Uncertainty in Obsessive-Compulsive Disorder." *Journal of Anxiety Disorders* 17 (2): 233–42.

Vaghi, M. M., Luyckx, F., Sule, A., Fineberg, N. A., Robbins, T. W., & De Martino, B. (2017). Compulsivity Reveals a Novel Dissociation between Action and Confidence. Neuron, 96(2), 348–354.e4.

Wang, Xiao-Jing. 2002. "Probabilistic Decision Making by Slow Reverberation in Cortical Circuits." *Neuron* 36 (5): 955–68.

Wang, Xiao-Jing, and John H Krystal. 2014. "Computational Psychiatry." *Neuron* 84 (3): 638–54.

Wong, Kong-Fatt, and Xiao-Jing Wang. 2006. "A Recurrent Network Mechanism of Time Integration in Perceptual Decisions." *Journal of Neuroscience* 26 (4): 1314–28.

Yang, Genevieve J, John D Murray, Grega Repovs, Michael W Cole, Aleksandar Savic, Matthew F Glasser, Christopher Pittenger, et al. 2014. "Altered Global Brain Signal in Schizophrenia." *Proceedings of the National Academy of Sciences* 111 (20): 7438–43.

Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The neural basis of error detection: conflict monitoring and the error-related negativity. Psychological review, 111(4), 931–959.

Zung, William WK. 1965. "A Self-Rating Depression Scale." *Archives of General Psychiatry* 12 (1): 63–70.