PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

# Tuned inhibition in perceptual decision-making circuits can explain seemingly suboptimal confidence behavior

**Brian Maniscalco**[1,2¤*], **Brian Odegaard**[3,4], **Piercesare Grimaldi**[5], **Seong Hah Cho**[6], **Michele A. Basso**[5,7,8], **Hakwan Lau**[4,6,8,9], **Megan A. K. Peters**[1,2,4,10,11¤*]

**1** Department of Cognitive Sciences, University of California Irvine, Irvine, California, United States of America, **2** Department of Bioengineering, University of California Riverside, Riverside, California, United States of America, **3** Department of Psychology, University of Florida, Gainesville, Florida, United States of America, **4** Department of Psychology, University of California Los Angeles, Los Angeles, California, United States of America, **5** Fuster Laboratory of Cognitive Neuroscience, Department of Psychiatry and Biobehavioral Sciences, University of California Los Angeles, Los Angeles, California, United States of America, **6** Department of Psychology, University of Hong Kong, Pokfulam, Hong Kong SAR, **7** The Jane and Terry Semel Institute for Neuroscience and Human Behavior, University of California Los Angeles, Los Angeles, California, United States of America, **8** Brain Research Institute, University of California Los Angeles, Los Angeles, California, United States of America, **9** State Key Laboratory of Brain and Cognitive Sciences, University of Hong Kong, Pokfulam, Hong Kong, SAR, **10** Interdepartmental Graduate Program in Neuroscience, University of California Riverside, Riverside, California, United States of America, **11** Department of Psychology, University of California Riverside, Riverside, California, United States of America

¤ Current address: Department of Cognitive Sciences, University of California Irvine, Irvine, California, United States of America
* bmanisca@uci.edu (BM); megan.peters@uci.edu (MAKP)

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

**Editor:** Jeff Beck, Duke University, UNITED STATES

**Received:** September 17, 2019

**Accepted:** February 8, 2021

**Published:** March 29, 2021

**Data Availability Statement:** Code for the Differential Tuned Inhibition model simulations is available at https://github.com/CNClaboratory/Tuned_Inhibition_PLOS_Comp_Bio_2021. Data for the monkey electrophysiology results described in S4 Text are available on Dryad (datadryad.org, doi:10.7280/D1T38T).

**Funding:** This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-20-1-0106 (to MAKP), and partially supported by the National Institutes of Health (nih.gov) (R01NS088628¬01 to

## Abstract

Current dominant views hold that perceptual confidence reflects the probability that a decision is correct. Although these views have enjoyed some empirical support, recent behavioral results indicate that confidence and the probability of being correct can be dissociated. An alternative hypothesis suggests that confidence instead reflects the magnitude of evidence in favor of a decision while being relatively insensitive to the evidence opposing the decision. We considered how this alternative hypothesis might be biologically instantiated by developing a simple neural network model incorporating a known property of sensory neurons: tuned inhibition. The key idea of the model is that the level of inhibition that each accumulator unit receives from units with the opposite tuning preference, i.e. its inhibition 'tuning', dictates its contribution to perceptual decisions versus confidence judgments, such that units with higher tuned inhibition (computing relative evidence for different perceptual interpretations) determine perceptual discrimination decisions, and units with lower tuned inhibition (computing absolute evidence) determine confidence. We demonstrate that this biologically plausible model can account for several counterintuitive findings reported in the literature where confidence and decision accuracy dissociate. By comparing model fits, we further demonstrate that a full complement of behavioral data across several previously published experimental results—including accuracy, reaction time, mean confidence, and metacognitive sensitivity—is best accounted for when confidence is computed from units without, rather than units with, tuned inhibition. Finally, we discuss predictions of our results

HL; R01EY13692 to MAB), and the Canadian Institute for Advanced Research Azrieli Global Scholars Program (to MAKP). The funders had no role in 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.

and model for future neurobiological studies. These findings suggest that the brain has developed and implements this alternative, heuristic theory of perceptual confidence computation by relying on the diversity of neural resources available.

## Author summary

The dominant view of perceptual confidence proposes that confidence optimally reflects the probability that a decision is correct. But recent empirical evidence suggests that perceptual confidence exhibits a suboptimal 'confirmation bias', just as in human decision-making in general. We tested how this 'bias' might be neurally implemented by building a biologically plausible neural network model, and showed that the 'bias' emerges when each neuron's degree of inhibition received from neurons with opposing tuning preferences dictates how it drives decisions versus confidence judgments. We additionally showed that alternative models lacking this architecture fail to capture known behavioral effects. These results challenge the dominant model, suggesting that the brain instead capitalizes on the diversity of available machinery (i.e., neuronal resources) to implement *heuristic*—not optimal—strategies to compute subjective confidence.

## 1. Introduction

A dominant idea in the study of perceptual decision-making is that confidence judgments optimally reflect the probability that a decision is correct [1–5]. Several models specifically stipulate that confidence is calculated via implementation of a diffusion framework: a decision is made when evidence for a decision reaches a certain threshold, and confidence reflects an optimal readout of the same information [3–9].

While this optimal 'probability correct' account of confidence has enjoyed significant empirical support, it seems difficult for it to account for cases where task performance and confidence dissociate [10–18]. Seemingly suboptimal behaviors have also been observed in post-decisional perceptual judgments other than confidence [19,20], leading these authors to hypothesize that these suboptimalities may stem from limitations on computational (i.e., neural) resources or a drive towards self-consistent behavior. One alternative theory of confidence, therefore, proposes that subjective confidence relies primarily on the magnitude of evidence supporting an observer's decision, while ignoring or downplaying evidence supporting alternative, unchosen decisions [10,14,16,18,21,22]. In other words, to compute confidence the system uses a suboptimal heuristic that overly relies on decision-congruent evidence magnitude rather than optimal computations. Indeed, a recent study reported evidence for these decision-congruent evidence confidence computations using human intracranial electrocorticography [23].

However, to date no biologically plausible mechanism has been proposed that might explain these dissociations between confidence and performance, or the decision-congruent confidence computations on which they seem to depend. We therefore developed a simple dynamic evidence accumulation network model to test a new hypothesis of how these computations might be implemented. This model extends previous work to incorporate a known property of perceptual circuitry: *tuned normalization* [24–26], meaning each neuron is characterized by the specific degree to which it is normalized by surrounding or nearby network activity [27,28], and specifically by units with opposing tuning preferences. In the present work, we use the more general term *tuned inhibition* to refer to any neural dynamics in which neurons are preferentially inhibited by other neurons with opposing tuning preferences, regardless of whether such dynamics are divisive in nature (as normalization processes are typically characterized to be), since our primary hypothesis concerns the general phenomenon of tuned inhibition without a particular concern for whether such inhibition is divisive or not. We hypothesized that each neuron's degree of tuned inhibition dictates how it differentially participates in discrimination decisions versus confidence judgments. Specifically, we reasoned that strongly inhibited 'differencing' neurons encode the balance of evidence for various perceptual interpretations (e.g. net accumulated evidence for leftwards or rightwards motion direction), and thus are reasonable candidates for making discrimination judgments. By contrast, less inhibited evidence accumulation neurons encode total overall evidence in favor of one perceptual interpretation (e.g. leftward motion) while ignoring evidence for alternative interpretations (e.g. rightward motion), and thus are reasonable candidates for implementing decision-congruent confidence computations. Therefore, the simple design principle that more inhibited differencing neurons drive decisions and less inhibited accumulator neurons drive confidence may be sufficient to account for some of the most counterintuitive empirical findings on confidence in perceptual decision-making.

We tested key predictions of a *Differential Tuned Inhibition* model instantiating this hypothesis using computational modeling, with exploratory supplemental results from single neuron physiology. The computational model simulations show strong support for our hypothesis: the model reproduces multiple empirical findings when confidence is computed primarily from less inhibited 'absolute evidence' units, but not when computed primarily from more inhibited 'differencing' units. With this approach, we extend previous conceptual insights to include dynamical evidence accumulation and reaction time effects, while also unifying multiple empirical datasets across multiple paradigms with a single coherent approach. We also show that meta-d', a measure of metacognitive sensitivity [29], provides a crucial target for model comparisons of perceptual confidence. Our results suggest that tuned inhibition may play a crucial role in how the brain differentially computes perceptual decisions and subjective confidence–thus revealing an important psychological function of this neuronal property.

## 2. Methods

### 2.1. Ethics statement

In S4 Text, we describe exploratory behavioral and electrophysiology results from the superior colliculus for one male rhesus monkey. Details of the surgical procedures used to implant electrodes are also provided in S4 Text. All experimental protocols were approved by the UCLA Chancellor's Animal Research Committee (IACUC, protocol number 2012–043) and complied with and generally exceeded standards set by the Public Health Service policy on the humane care and use of laboratory animals.

### 2.2. The *Differential Tuned Inhibition* model

The model presented here is a dynamic evidence accumulation network with tuned inhibition, where units with different inhibition tuning differentially contribute to perceptual decisions and confidence.

To investigate how decision-congruent evidence might be biologically implemented, we began by considering known properties of perceptual decision-making circuitry. It is well known that normalization is a canonical neural computation throughout the cortex [27,28,30–34]. Further, it was recently reported that neurons in primary sensory areas exhibit *tuned* *normalization* (here called *tuned inhibition* for generality), i.e. that each neuron possesses a unique, consistent degree of received inhibition: some neurons are very sensitive to activity of other units in the network (especially those which have different tuning preferences), while others operate more independently [24–26]. Previous implementations of evidence accumulation models for perceptual decisions have typically considered how a single level of normalization or inhibition—in a single set of units that drive both perceptual decisions and confidence —can account for behavioral data [3–8,21,35]. However, we now know that a range of inhibition tuning exists, at least in sensory cortices. We hypothesized that these neuron-by-neuron variations in inhibition may reflect not noise or measurement error, but meaningful properties of the perceptual decision-making circuitry [24–26]. We refer to this model as the *Differential Tuned Inhibition* model, or the *Tuned Inhibition* model for short.

But *how* might tuned inhibition be utilized in a behaving organism? To answer this, we should consider the tasks an organism must successfully execute in an ecologically valid environment. The ability to discriminate among multiple possible stimulus identities is certainly important, and for this type of task an observer ought to rely on a system that is able to average out noise, i.e. is less susceptible to random fluctuations in signal. Thus, for these discrimination-type tasks, a strong degree of inhibition would be desirable, as it has been shown that neurons with stronger tuned inhibition do exhibit weaker pairwise correlations [25,26]. But it is equally important that an organism also be able to detect a stimulus in the first place, regardless of its identity. For these detection-type tasks, such strong inhibition would actually be undesirable, as minute evidence amounts may be informative; therefore, weakly inhibited neurons ought to play a stronger role in detection-type tasks. As both of these task types are critical for an organism's survival, it seems unlikely that a system would only be optimized for one or the other, which could in theory explain the presence of tuned inhibition.

In light of this discussion, and of the empirically observed tuned inhibition in cortical areas, a biologically plausible model of sensory evidence accumulation ought to implement more than one level of inhibition and consider how such variations in inhibition tuning may affect a neuron's role in the circuitry. Further, such stratification of tuned inhibition could provide a neural mechanism to explain findings that confidence judgments rely on the magnitude of decision-congruent evidence [10,14,17,18,21–23]. Specifically, the output of less inhibited 'detection' neurons could be used to index decision-congruent evidence and therefore be used for confidence rating. This suggests that inhibition tuning provides a biologically plausible mechanism to keep track of decision-congruent evidence independently of evidence favoring other possible choices by relying on the less inhibited portion of the circuitry, while allowing the system to still capitalize on the beneficial consequences of inhibition by opposingly-tuned neurons by relying on the more inhibited portion when discriminating among possible stimulus identities. We therefore hypothesized that inhibition tuning might specifically dictate a neuron's contribution to discrimination versus confidence judgments in decision-making circuitry.

We examined this hypothesis by incorporating tuned inhibition [24,25] into a dynamic evidence accumulation network (Fig 1). Intuitively, this network's architecture can be summarized as follows. Accumulator units tuned to varying stimulus alternatives accumulate momentary stimulus evidence. Downstream 'differencing' units receive excitatory and inhibitory input from accumulator units having opposing tuning preferences, effectively performing a subtraction to yield the balance of evidence favoring one stimulus alternative over the other. A discrimination decision is made when a differencing unit with a given tuning preference reaches a threshold level of evidence.

Following the perceptual decision, additional evidence accumulation occurs in order to form a confidence judgment [9]. Confidence is then evaluated by comparing accumulator unit

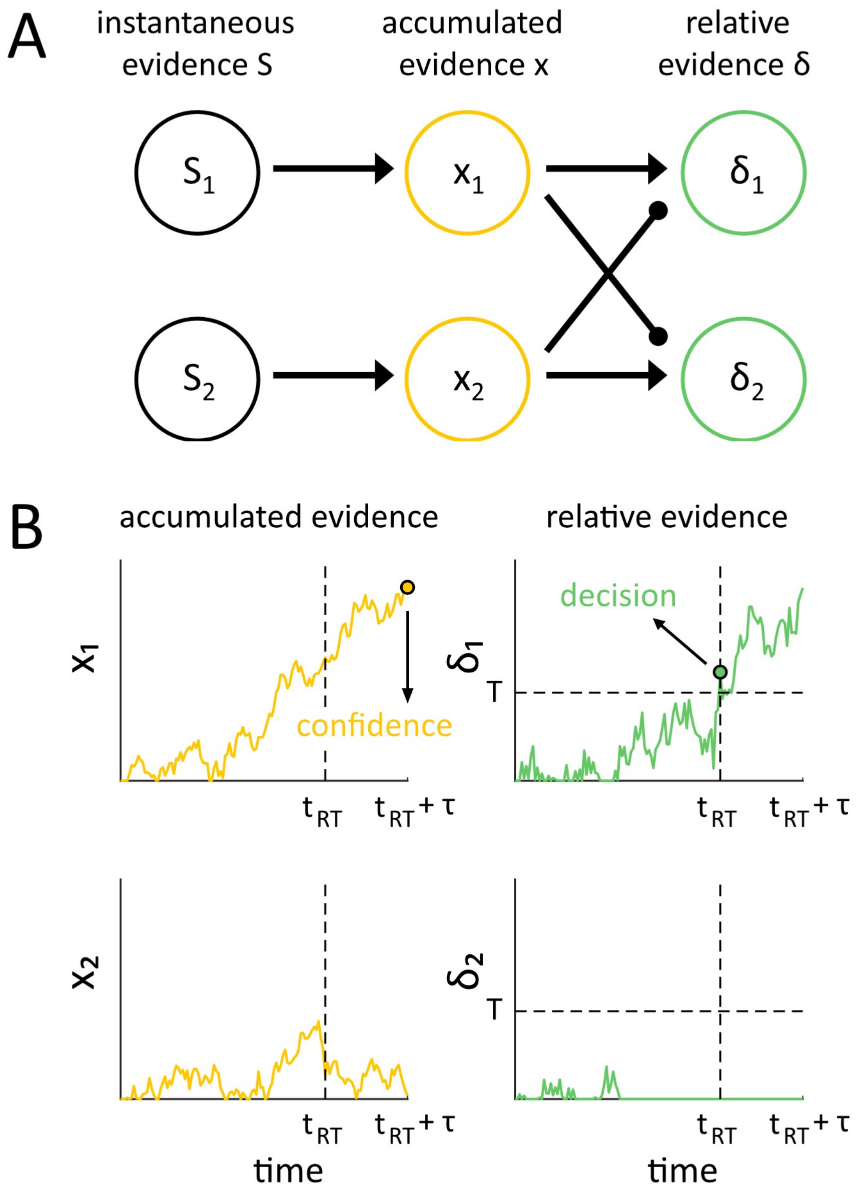

**Fig 1. The Differential Tuned Inhibition model diagram and sample activity traces.** (A) In the 2-stage model, instantaneous evidence from a source $S_i$ is accumulated by independent accumulators $x_i$ tuned to each stimulus type *i*. Differencing units $\delta_i$ then compute the difference in accumulated evidence for each stimulus alternative, implemented through feed-forward excitation from units with the same preferences and inhibition from accumulator units with opposing preferences. (B) A decision is made by the model when the activity of one of the differencing units hits a bound, i.e. when enough relative evidence for one stimulus over another has been accumulated. Confidence in the main Differential Tuned Inhibition model is then read out after delay $\tau$ from the independent accumulator $x_D$ corresponding to the decision D that was made. In an alternative model, we also tested how this framework might perform if confidence were read out from the differencing unit $\delta_D$; see Methods for details.

https://doi.org/10.1371/journal.pcbi.1008779.g001

activity for the chosen stimulus alternative to a set of decision thresholds for rating confidence. The duration τ of this post-decision evidence accumulation controls metacognitive sensitivity, i.e. the efficacy with which confidence ratings discriminate correct from incorrect responses, as measured e.g. by the signal detection theory measure meta-d' [29]. In general, longer post-decision accumulation periods yield higher metacognitive sensitivity (consistent with empirical patterns reported in [36]). When τ = 0 and confidence is determined at the same instant as the perceptual decision, there is very little variability in perceptual evidence values (since, by definition, the perceptual decision is made when perceptual evidence in the differencing units arrives at a fixed threshold value), which has the consequence that meta-d' ≈ 0 (see e.g. S1C and S1D Fig), meaning confidence is at chance levels of discriminating correct from incorrect responses. Thus, incorporating a post-decision evidence accumulation stage with τ > 0 in this kind of evidence accumulation model is important to capture empirical patterns of metacognitive sensitivity.

For simplicity, we assume that evidence accumulation rates remain constant in the post-decision stage. This assumption need not imply that the physical stimulus remains available to perception for the duration of the decision process, as evidence accumulation could depend in part on internal processes that continue even after stimulus offset. Consistent with this idea, prior work has demonstrated that for briefly presented and masked stimuli, diffusion models assuming constant drift rate fit the data better than ones assuming drift rates that vary with stimulus duration and mask onset, suggesting that evidence accumulation rates can persist over time even after stimulus offset [37].

Crucially, the roles of these unit types in the perceptual decision making process depends on their level of inhibition tuning, and this weighting differs for discrimination decisions and confidence ratings. More inhibited 'differencing' units determine discrimination decisions, since they effectively encode the accumulated balance of evidence for one stimulus alternative versus the others by virtue of their receipt of inhibition from units with opposing tuning preferences. By contrast, less inhibited accumulator units determine perceptual confidence, since they effectively encode a "pure" representation of the raw magnitude of independent evidence supporting each decision alternative, regardless of evidence supporting other possible decisions. Details of model implementation and all simulations follow in the next sections.

## 2.3. Model specifics

The Differential Tuned Inhibition model consists of two main stages of processing: (1) an initial evidence accumulation stage, in which independent sensory evidence for different perceptual interpretations is independently integrated over time, and (2) a subsequent evidence comparison stage, in which the independently-accumulated evidence for perceptual alternatives in the previous stage is directly compared by a subtractive (differencing) process which could be biologically implemented through feed-forward inhibition (Fig 1).

To capture perceptual dynamics in experiments in which an observer must use noisy sensory evidence to decide between two perceptual alternatives (e.g. deciding if a grating is tilted left or right, if dot motion is moving leftwards or rightwards, etc.), we model two independent accumulator units (which can be thought of as corresponding to two independent pools of unnormalized accumulator neurons) corresponding to each perceptual alternative, where each evolves according to the equation:

$$x_i(t) = max(x_i(t-1) + S_i + \varepsilon_i(t), 0) \tag{1}$$

In this equation, i denotes stimulus alternative, $S_i$ denotes instantaneous evidence for stimulus i (assumed to be constant over time), and $\varepsilon_i(t)$ is a noise term such that $\varepsilon_i(t) \sim N(0, \sigma)$. The max operation rectifies the accumulator unit to ensure non-negative values.

At the second stage, the balance of evidence supporting stimulus i relative to stimulus j is computed as the instantaneous difference between the evidence accumulated thus far for i and j:

$$\delta_i(t) = max(x_i(t) - x_j(t) + \xi_i(t), 0) \tag{2}$$

where ξ$_i$(t) is a noise term such that ξ$_i$(t) ~ N(0, σ). (For simplicity, we assume that noise terms at each level of processing, ε$_i$(t) and ξ$_i$(t), are drawn from the same distribution.) Once again, the results are rectified to ensure non-negative values. Evidence accumulation proceeds until one of the strongly inhibited 'differencing' units δ$_i$ achieves some threshold value T, at which point the observer decides upon perceptual interpretation i as their decision D for this trial:

$$D = \begin{cases} i \ if \ \delta_i(t) > T \\ j \ if \ \delta_j(t) > T \end{cases} \tag{3}$$

Reaction time $t_{RT}$ is considered to correspond to the value of t at which $\delta_D$ first surpasses T.

After the initial perceptual decision is made, evidence accumulation in the uninhibited units continues for a fixed number of time steps τ, following previous convention [9]. Once τ time steps have passed, confidence is read out as the value of evidence in the uninhibited accumulator unit corresponding to the perceptual decision, i.e.

$$C_x = x_D(t_{RT} + \tau) \tag{4}$$

Thus, whereas the perceptual decision D depends on the balance of evidence between stimulus alternatives $\delta_D$, confidence $C_x$ depends on the total amount of "uninhibited" or "absolute" accumulated evidence in favor of the perceptual decision, $x_D$. This divergence in the computation of perceptual decision and confidence embodies the decision-congruent evidence effect discovered in empirical studies and captures the central phenomenon of interest for this manuscript.

To directly test the necessity of this model architecture for capturing the behavioral effects reported in the literature, we also implement an alternative control model consistent with dominant theory that decisions and confidence judgments are computed from the same units, i.e.

$$C_\delta = \delta_D(t_{RT} + \tau) \tag{5}$$

This alternative model is logically consistent with canonical models for computing confidence from the accumulated balance of evidence (e.g., [3–5]).

It is important at this point to clarify a point of potential misunderstanding. Note that in both Eq 4 and Eq 5, confidence is read out from a decision unit whose tuning preference matches the perceptual decision (denoted by the D subscripts in both equations). Thus, on one possible reading of the term "decision congruent evidence," one might conclude that confidence depends on decision congruent evidence in both Eqs 4 and 5, since in both cases, confidence is read out from a unit whose stimulus tuning preference matches the perceptual decision. However, when we use the term "decision congruent evidence" in this manuscript, consistent with prior usage in the literature, we mean to refer to evidence that supports the perceptual decision one has chosen, independent of evidence for alternative choices. In this sense of the term, only Eq 4 qualifies as computing confidence from decision-congruent evidence, since only in this case is confidence insensitive to evidence for the alternative, unchosen perceptual interpretation(s).

For both the main Tuned Inhibition $C_x$ model and the alternative $C_\delta$ model, we compare model outputs to empirical confidence data by converting raw confidence C values to an ordinal rating scale value R by comparing the respective Cs to a series of confidence threshold values $U_r$, as follows:

$$R = 1 + \sum_{r=1}^{N_r - 1} (C > U_r) \tag{6}$$

where $N_r$ is the number of rating scale options available (e.g. $N_r = 4$ on a 4-point confidence scale), and ($C > U_r$) is a logical comparison evaluating to 1 if the inequality is true and 0 otherwise. Thus, R is a simple count of how many of the confidence thresholds $U_r$ are surpassed by C, with the constant 1 added to set 1 as the minimum confidence rating value.

## 2.4. Simulations

We tested the model by assessing its ability to capture empirical dissociations between perceptual task performance and confidence in three representative data sets: a dissociation between d' and meta-d' (the trial-by-trial correspondence between accuracy and confidence [29]) observed in Maniscalco, Peters, & Lau [14], and dissociations between d' and mean confidence observed in Experiments 1A and 2B of Koizumi, Maniscalco, & Lau [10]. Details of the experimental designs and simulation results are discussed below.

For each data set, we fit the Differential Tuned Inhibition model (with confidence computed from absolute evidence accumulators) and the alternative model (with confidence computed from the difference between accumulators) to the data and compared the ability of the models to capture the relevant dissociations. In connection with Eqs 4 and 5, we call these the $C_x$ model and $C_\delta$ model for short, respectively. We expected that the $C_x$ model would outperform the $C_\delta$ model in its ability to capture the dissociations, thus lending further support to the hypothesis that confidence depends primarily on decision-congruent evidence.

We followed a similar approach in fitting the models to these diverse data sets, as outlined below, though certain details of model fitting were particular to each data set. Full details for model fitting are provided in S1 Text.

**2.4.1. Selecting values for σ and T.** First, we arbitrarily set accumulation noise $\sigma = 0.1$. The choice of this parameter value is arbitrary since, if no parameter values are fixed, identical simulation results can be obtained by a simple scaling of the model parameters. We then set decision threshold $T = 1$ to ensure that, even in the absence of stimulus drive S (i.e. $S_1 = S_2 = 0$), accumulation of noise alone could reach T within a reasonable number of time steps, while still ensuring that at least several time steps must pass for this to occur (in 10 repetitions of simulations with 10,000 trials each, average median RT = 80.6 and average minimum RT = 7.3). These choices for σ and T formed a fixed reference against which other parameters of the model could be optimized. We found that similar simulation results occurred when using different values for T, which can be readily verified using the simulation code available online.

**2.4.2. Fitting S to d'.** Next, for each data point in the data set, we found what value of stimulus drive S would be required to yield the desired value of d'. We did this by performing simulations at 10 evenly spaced values of S and fitting a quadratic polynomial to the resulting d' vs S curve, which provided an excellent fit across a range of d' values from ~ 0–3 (S1A and S1B Fig). Using the fitted polynomial equation, we could solve for what value of S yielded the value of d' to be fitted.

**2.4.3. Fitting $U_r$ to confidence probability distributions and τ to meta-d'.** The parameter τ determines how many time steps pass after the initial perceptual decision is made, during which evidence continues to accumulate before a confidence judgment is formed. Larger values of τ correspond to more evidence accumulation prior to confidence rating and therefore higher values of meta-d' (i.e., confidence ratings that carry more information about task accuracy). Thus, empirically observed values of meta-d' can serve as a guide for appropriate values of τ.

In each data set simulation, we fit τ to the meta-d' value of only one representative data point, and used this value of τ for all subsequent simulations of that data set. This approach ensured that τ was held constant across all conditions. Importantly, as a consequence, the fitting procedure guaranteed a close fit to meta-d' in only one data point, and simulated meta-d' values at all other data points were unconstrained by further parameter fitting and instead arose as a consequence of the value of τ fitted to the representative meta-d' value.

To accomplish the fit, we performed simulations at 10 evenly spaced values of τ and fitted a quadratic polynomial to the resulting meta-d' vs τ curve, which provided an excellent fit across a range of meta-d' values from ~ 0 –d' (S1C and S1D Fig). Using the fitted polynomial equation, we could solve for what value of τ yielded the value of meta-d' to be fitted for the single fitted data point.

In order to compute meta-d', continuous confidence values ($C_x$ or $C_\delta$, depending on the model being used) first had to be converted to a 4-point rating scale (corresponding to the 4-point confidence rating scale used in all three data sets to be fitted), which in turn required specifying the values of the confidence thresholds $U_r$. For each simulation, we set $U_r$ such that the probability distribution of simulated confidence ratings across all trials of all conditions exactly matched the corresponding empirical probability distribution. More formally, we computed $U_r$ as

$$U_r = quantile(C, \sum_{i=1}^{r} P_{data}(conf = i)) \tag{7}$$

where C corresponds to $C_x$ or $C_\delta$, depending on the model being used, quantile(C, p) returns the quantile of the distribution C corresponding to the cumulative probability p, and $P_{data}(conf = i)$ is the empirical probability distribution of confidence ratings.

As noted above, full details of the model fitting procedures are provided in S1 Text.

## 3. Results

### 3.1. Simulating data from Maniscalco, Peters, & Lau 2016 [14]

Maniscalco, Peters, & Lau [14] used signal detection theory modeling to predict that, if human observers do indeed use decision-congruent evidence to judge confidence, then under certain conditions it should be possible to observe a counterintuitive dissociation whereby metacognitive sensitivity *decreases* even as perceptual task performance *increases*. They conducted an experiment to test the prediction and verified that human observers did indeed exhibit this surprising dissociation (Fig 4 of [14]; results reproduced in Fig 2), providing strong support for the decision-congruent evidence model of confidence.

The experiment used a simple 2IFC task in which circular noise patches were presented to the left and right of fixation on every trial, with a grating embedded in the left or right noise patch. Subjects had to report which side the grating was on and then rate confidence on a 4-point scale. Crucially, the contrast of gratings presented on one side of the screen was held constant (call it $C_{S1}$), while the contrast of gratings presented on the other side could take on one of 5 possible values $C_{S2}$. $C_{S2}$ spanned a range of values below and above $C_{S1}$, with the intermediate value of $C_{S2}$ equal to $C_{S1}$. Briefly, the logic behind the predicted dissociation is as follows: If subjects rate confidence according to decision-congruent evidence, then it follows from a two-dimensional signal detection theory model that incorrect "S1" responses should yield higher mean confidence as $C_{S2}$ increases. Since $C_{S1}$ is constant, correct "S1" responses have a fixed level of mean confidence. Thus, the net effect of increasing $C_{S2}$ is to increase task performance (d') while simultaneously increasing confidence in incorrect "S1" responses but

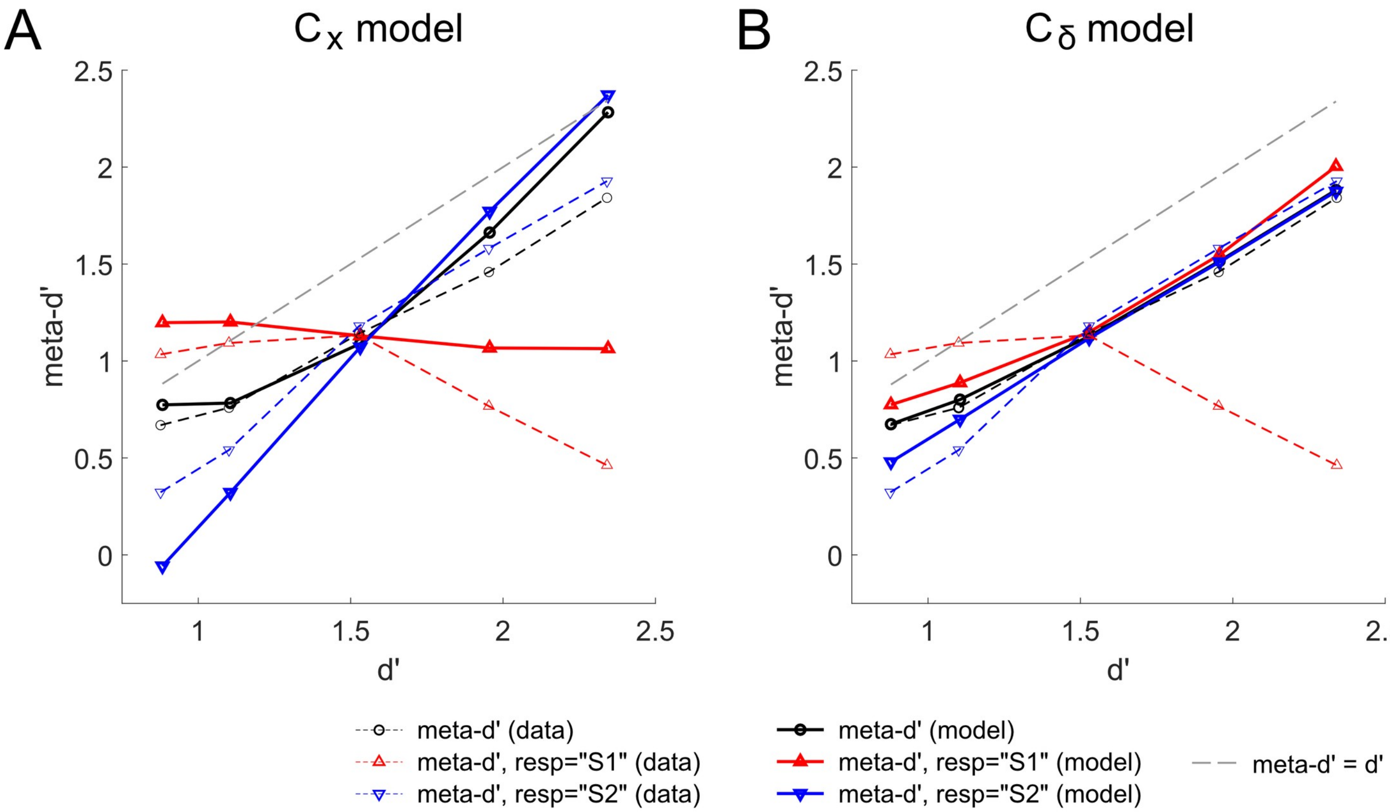

**Fig 2. Modeling results for the effects reported by Maniscalco and colleagues [14].** The fitted main $C_x$ model (A) reproduces the X-shaped crossover in response-conditional meta-d' that was observed in the empirical data: as d' increases, response-conditional meta-d' for "S2" responses increases whereas meta-d' for "S1" responses decreases. In contrast, such an X-shaped crossover in response-conditional meta-d' is not even qualitatively present for the alternative $C_\delta$ model (B).

https://doi.org/10.1371/journal.pcbi.1008779.g002

not correct "S1" responses, leading to a decrease in metacognitive sensitivity (meta-d') for "S1" responses. This predicted pattern in confidence for incorrect "S1" responses is indeed what was observed (S3 Fig of [14]; results reproduced in S2A Fig). For further discussion, see S2 Text.

To model the data, we structured the model simulations to mirror the experimental design feature whereby one stimulus ("A," corresponding e.g. to gratings presented on the left side of fixation) had constant stimulus strength $S_A$ across all simulated trials, whereas the other stimulus ("B," corresponding e.g. to gratings presented on the right side) could take on one of five possible values $S_{B,i}$, with the intermediate value $S_{B,3}$ defined to be equal to $S_A$. The parameter controlling meta-d' (τ) was fit to overall meta-d' at the intermediate level of task performance where stimulus strength for A and B was equal and the empirical response-conditional meta-d' curves intersected (Fig 2A). Full details of model fitting are provided in S1 Text.

In good agreement with the empirical data, simulated response-conditional meta-d' curves for the Differential Tuned Inhibition model exhibit an X-shaped cross-over effect, such that meta-d' for different response types either increases or decreases as d' increases. The range of meta-d' values exhibited by the model is both plausible and consistent with meta-d' found in other empirical studies as well. Notably, since τ was fit only to overall meta-d' at the single intermediate level of task performance, the meta-d' curves at all other data points were not constrained by an explicit fitting procedure but rather arise naturally as a consequence of the single set of fitted parameters.

Importantly, as described in Methods (Sections 2.3 and 2.4), we also tested the importance of computing confidence from uninhibited evidence accumulation to the data fit by repeating the above-described simulation procedures, but this time computing simulated confidence from $C_\delta$ rather than $C_x$, and using the fitted parameter $\tau_\delta$ rather than $\tau$. As expected, this alternative version of the model using $C_\delta$ was not able to capture, even qualitatively, the X-shaped cross-over dissociation in response-conditional meta-d' observed in the empirical data from Maniscalco and colleagues [14] (Fig 2B).

Further simulation results for this experiment are presented in S2 Text. There we compare empirical response-conditional mean confidence for correct and incorrect responses to the values produced in the $C_x$ and $C_\delta$ model simulations (S2 Fig), and show topographical plots of distributions of $C_x$ and $C_\delta$ to provide further intuition for why the former, and not the latter, can capture the response-conditional meta-d' dissociation (S3 Fig). We also point out conceptual and computational similarities between the $C_x$ model and the two-dimensional signal detection theory (2D-SDT) model used by Maniscalco et al. [14] as a theoretical frame of reference for their experimental design and empirical findings, arguing that these similarities are precisely what allow the $C_x$ model to capture the dissociation. Finally, in S3 Text and S4–S7 Figs, we demonstrate that a leaky competing accumulator (LCA) implementation [35] of the tuned inhibition model cannot capture the data as well as the model implementation explored in the main manuscript, and argue that the LCA model performs worse in part because it does not map as cleanly onto the 2D-SDT model structure as does the main model.

## 3.2 Simulating data from Koizumi, Maniscalco, & Lau 2015 [10] Experiment 1A

Koizumi, Maniscalco, & Lau [10] experimentally controlled stimulus properties to yield conditions where task performance (d') was similar, yet mean confidence differed. In their Experiment 1A, subjects performed a grating tilt discrimination task. However, stimuli were actually composed of two superimposed gratings tilting in opposite directions, one with higher contrast ("Positive Evidence" or "PE" for short; the correct decision) and one with lower contrast ("Negative Evidence" or "NE"; the incorrect decision). Subjects had to indicate whether the higher contrast grating was tilting left or right and then rate confidence on a 4-point scale.

The key experimental manipulation was the introduction of High PE and Low PE conditions, in which the contrast of the NE gratings was set to 0.7*(PE grating contrast) and 0.35*(PE grating contrast), respectively, and PE grating contrast was controlled by thresholding procedures to achieve a criterion level of task performance (d'). Due to the distractor NE stimulus being stronger in the High PE condition, a higher level of PE was required to achieve a given level of d', with the net effect that d' was similar in Low PE and High PE conditions, whereas PE (and NE) were higher in the High PE condition than in the Low PE condition. Additionally, there were two levels of task difficulty at each level of PE, thus giving rise to four experimental conditions in all: "High PE, Easy," High PE, Difficult," "Low PE, Easy," and Low PE, Difficult." Koizumi et al found that the High PE and Low PE conditions yielded similar task performance (d'), but mean confidence was higher for High PE stimuli (Fig 1D of [10]).

Similarly to Maniscalco et al 2016 [14], the dissociation effect in Experiment 1A of Koizumi et al 2015 [10] is thought to arise from observers using decision-congruent evidence to assess confidence. Thus, we similarly expected that the dissociation between d' and confidence would be captured well by simulations using $C_x$ (corresponding to computing confidence from absolute accumulated decision-congruent evidence), but not by simulations using the alternative $C_\delta$ (corresponding to computing confidence from the 'differencing' units, i.e. the balance of evidence).

To attain a comprehensive understanding of model behavior across all levels of low and high PE, we exhaustively mapped the behavior of the $C_x$ and $C_\delta$ models across different levels of PE and NE, as follows. Since all stimuli in Koizumi et al. [10] Experiment 1A featured non-zero levels of positive and negative evidence, we modeled each condition using two stimulus drive parameters $S_{PE}$ and $S_{NE}$. In all simulations, we set a constraint such that $S_{NE} = \alpha\ S_{PE}$ for $0 < \alpha < 1$, and further constrained $\alpha$ to be lower in the Low PE condition, i.e. $\alpha_{low\ PE} < \alpha_{high\ PE}$. We selected $\alpha_{low\ PE} = 0.1$, 0.3, and 0.5 as three representative values of $\alpha$ to explore for the Low PE condition. For each level of $\alpha_{low\ PE,\ i}$, we explored a corresponding set of values of $\alpha_{high\ PE,\ ij}$ where the $\alpha_{high\ PE,\ ij}$ values were set as equally spaced values between $\alpha_{low\ PE,\ i} + 0.1$ and 0.9, incrementing by steps of 0.1. For each pairing of $\alpha_{low\ PE,\ i}$ and $\alpha_{high\ PE,\ ij}$, we fit the model parameters to the average d' and meta-d' values in the Low PE, Easy and Low PE, Difficult conditions and observed how this influenced the difference in mean confidence between the simulated High PE and Low PE conditions (Fig 3A, 3E and 3I).

Simulations results revealed that the main $C_x$ model can capture a wider range of differences in confidence due to manipulation of PE level, which allows it to capture the observed magnitude of the (High PE confidence–Low PE confidence) effects across all analyzed choices of $\alpha_{low\ PE,\ i}$ using relatively small increments in the corresponding $\alpha_{high\ PE,\ ij}$ (for $\alpha_{low\ PE} = 0.1$, 0.3, and 0.5, the effect of PE level on confidence is captured by setting $\alpha_{high\ PE} = 0.2548$, 0.3966, and 0.5525, respectively; Fig 3A, 3E and 3I). By contrast, the alternative $C_\delta$ model predicts only a very small influence of PE level on mean confidence; at $\alpha_{low\ PE} = 0.1$ it can account for the confidence effect only by positing a massively stronger PE level at $\alpha_{high\ PE} = 0.8928$ (Fig 3A), and when $\alpha_{low\ PE} > 0.1$, no value of $\alpha_{high\ PE} \leq 0.9$ can account for the magnitude of the effect (Fig 3E–3I).

To provide a more comprehensive characterization of the model fits to the empirical data, we also compared empirical and simulated data for the differences in reaction time and meta-d' induced by the High PE and Low PE conditions. RT and meta-d' data were not reported in Koizumi et al [10]; here we re-analyzed their data set and report the RT and meta-d' data for the first time (Fig 4). In the empirical data, meta-d' increased and RT decreased with increasing d', but neither was appreciably modulated by PE level.

To assess the effect of PE level on simulated RT, we computed a relative RT effect size $\eta_{RT}$, defined as

$$\eta_{RT} = (RT_{high\ PE,med} - RT_{low\ PE,med})/(RT_{low\ PE,easy} - RT_{low\ PE,difficult}) \tag{8}$$

The subscripts “high PE, med” and “low PE, med” denote that these RT values were taken from simulations of the High PE and Low PE conditions, using $S_{low\ PE}$ and $S_{high\ PE}$ values chosen to match mean d' across difficulty levels in the Low PE condition. The subscripts “low PE, easy” and “low PE, difficult” denote that these RT values were taken from the simulations of the Low PE condition, using $S_{low\ PE}$ values chosen to match d' in the Low PE, Easy and Low PE, Difficult conditions. Thus, $\eta_{RT}$ measures simulated change in RT at an intermediate d' value due to High vs Low PE (the numerator), relative to simulated change in RT in the Low PE condition due to Easy vs Difficult stimulus settings (the denominator).

Simulation results for (High PE meta-d'–Low PE meta-d') and $\eta_{RT}$ are shown in Fig 3B, 3F and 3J and Fig 3C, 3G and 3K, respectively. In general, the $C_x$ and $C_\delta$ models make differing predictions about how PE level influences meta-d', whereas they make the same prediction about how PE level influences RT (since perceptual decision making is driven by the same mechanism in both models). Of central interest for our purposes, however, is how these models predict meta-d' and RT to behave at the specific values of $\alpha_{low\ PE}$ and $\alpha_{high\ PE}$ that yield differences in mean confidence matching those observed in the data. These points are marked by

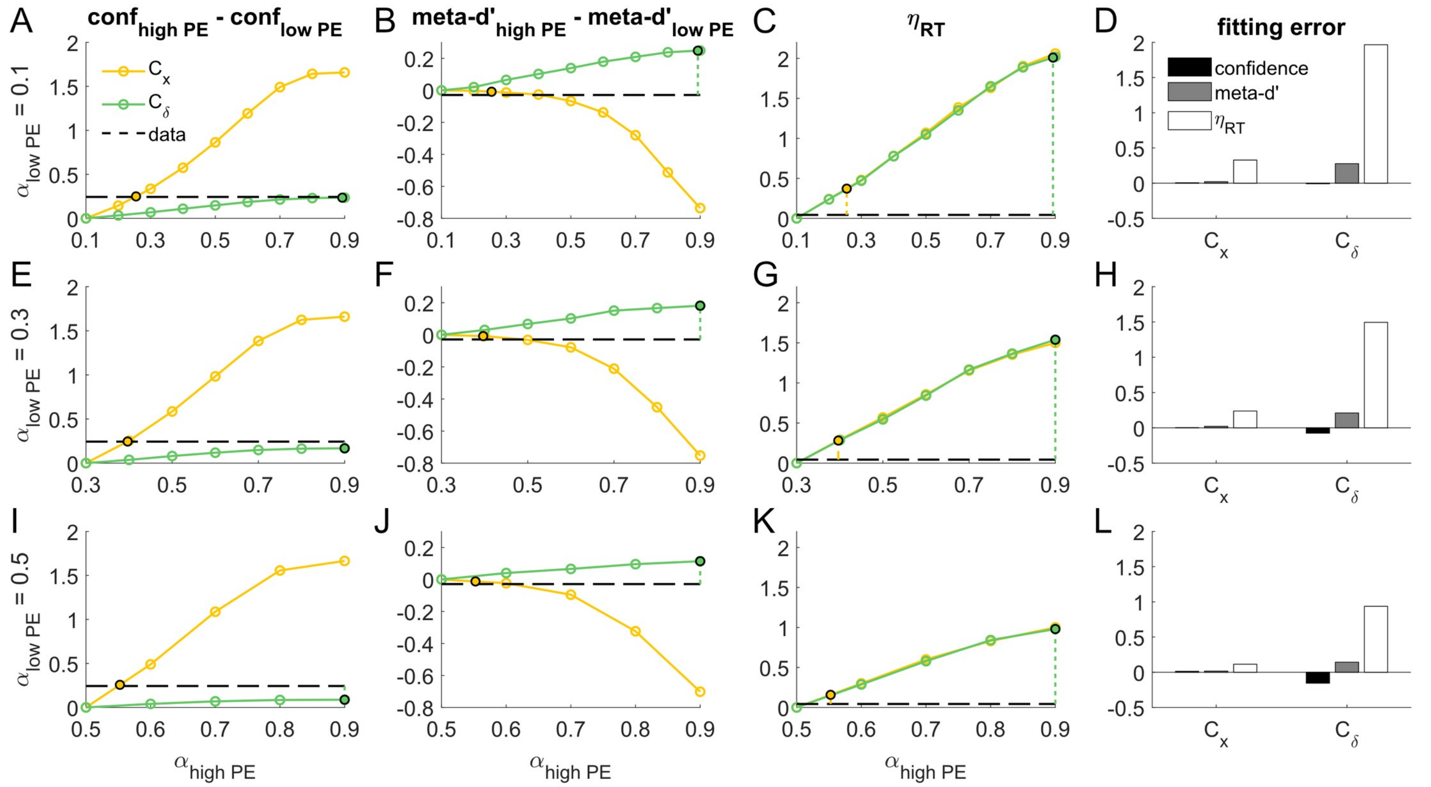

**Fig 3. Comprehensive simulations showing fitting procedures for the main $C_x$ model and alternative $C_\delta$ model for Koizumi et al.'s [10] Experiment 1A.** Rows show simulation results for different settings of $\alpha_{low\ PE}$, which determines the relative strength of positive evidence (PE) and negative evidence (NE) in simulations for the Low PE condition of Koizumi et al., such that $S_{NE} / S_{PE} = \alpha_{low\ PE}$. Columns show performance-matched confidence differences (i.e. differences in confidence when d' is exactly matched at the mean d' value of the low PE condition), performance-matched meta-d' differences, relative RT effect sizes ($\eta_{RT}$), and model fitting error due to PE manipulations for Koizumi et al.'s Experiment 1A. For each of the $C_x$ and $C_\delta$ models and each PE/NE balance in the Low PE condition ($\alpha_{low\ PE}$), we determined the level of NE in the High PE condition ($\alpha_{high\ PE}$) that best matched the observed magnitude of performance-matched confidence differences in the empirical data (dashed horizontal lines, column 1); these are marked with black circles. We then "read out" the predicted meta-d' differences and relative RT effect sizes (columns 2 and 3, respectively) for each of these best-fitting PE/NE levels to get a comprehensive picture of the behavior of the system for both the $C_x$ and $C_\delta$ models. Fitting error was computed for each model simply as the difference between the predicted value and the empirical value for confidence differences, meta-d' differences, and relative RT effect sizes (column 4). Across all PE/NE balance levels tested, the $C_x$ model produced comprehensively good fit to the empirical data from Koizumi et al.'s Experiment 1A; in contrast, the alternative $C_\delta$ model must be pushed to extremely high levels of $\alpha_{high\ PE}$ to capture the empirically observed confidence differences due to PE manipulations, leading to substantial errors in the corresponding fits for meta-d' and RT.

https://doi.org/10.1371/journal.pcbi.1008779.g003

filled circles and vertical dashed lines in Fig 3. For purposes of comparison to simulated data, we used linear fits to interpolate the empirical meta-d' and RT values at d' = 1.71 (corresponding to the average d' across the Low PE, Easy and Low PE, Difficult conditions) and computed the difference in the interpolated values for the High PE and Low PE conditions, yielding empirical values of High PE meta-d'–Low PE meta-d' = -0.029 and $\eta_{RT}$ = 0.044 (horizontal dashed lines in Fig 3).

At levels of $\alpha_{low\ PE}$ and $\alpha_{high\ PE}$ that yield a perfect fit to the effect of PE level on confidence, the main $C_x$ model also exhibits an accurate fit to the effect of PE level on meta-d', correctly predicting that PE level has a negligible effect on meta-d' (PE effect on meta-d' = -0.029 in the data; -0.008, -0.008, and -0.012 in the model for $\alpha_{low\ PE}$ = 0.1, 0.3, and 0.5 respectively). The $C_x$ model incorrectly predicts that RT should be faster under High PE than Low PE, but the predicted magnitude of this effect is modest compared to the speed-up in RT due to Easy vs Difficult stimulus settings in the Low PE condition ($\eta_{RT}$ = 0.044 in the data; 0.372, 0.281, 0.158 in the model for $\alpha_{low\ PE}$ = 0.1, 0.3, and 0.5 respectively). In the empirical data ($RT_{low\ PE,\ easy}$—

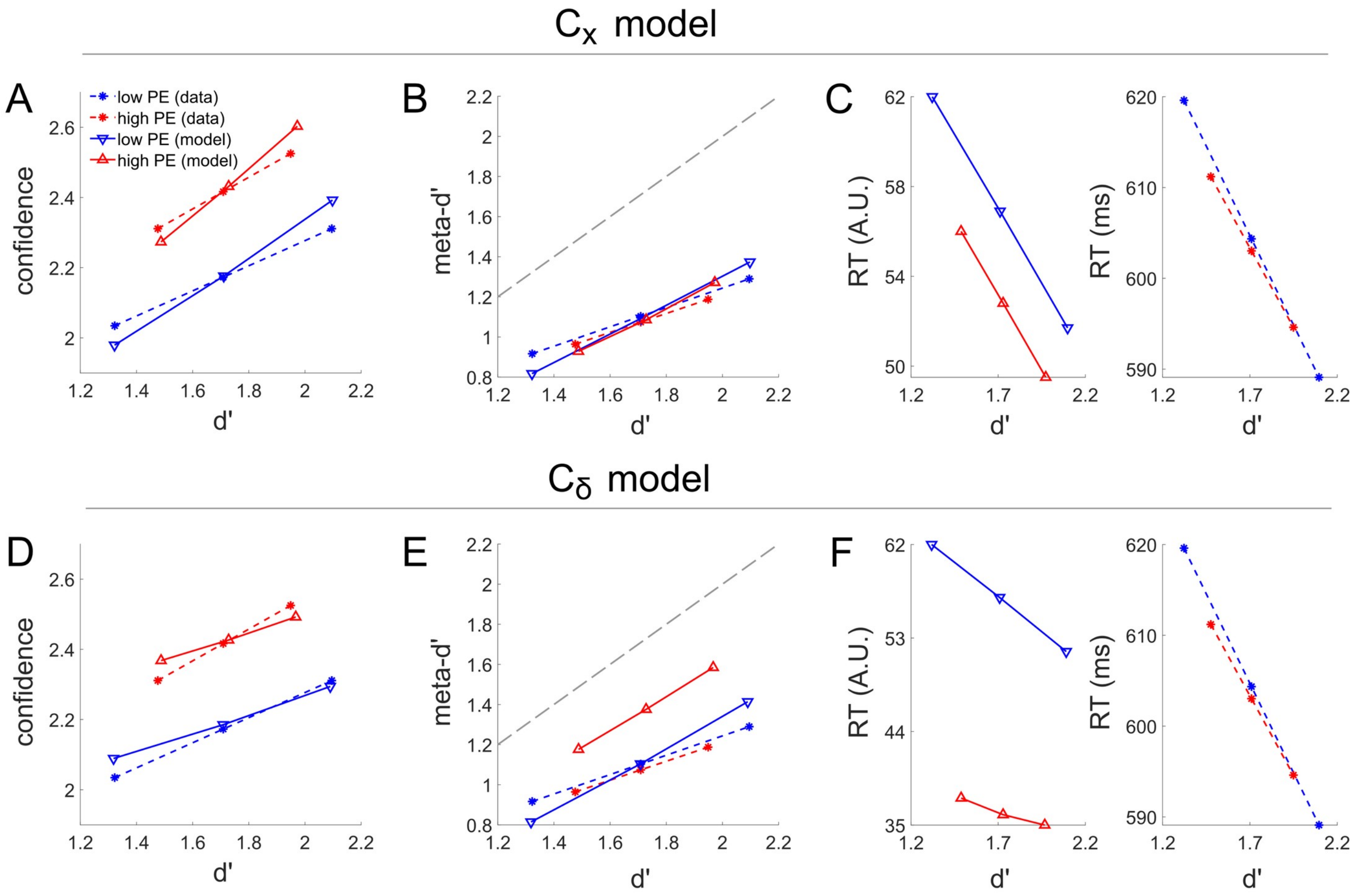

**Fig 4. Modeling results for the effects reported by Koizumi and colleagues [10] in their Experiment 1A with $\alpha_{\text{low PE}} = 0.1$ (corresponding to the first row of Fig 3) and parameter values chosen to best capture the effect of PE level on mean confidence (corresponding to the data points shown in black circles in Fig 3).** The fitted main $C_x$ model reproduces the differential confidence for matched d' performance for high versus low PE stimuli (A), as well as the overlapping distributions of meta-d' as a function of d' (B). The main $C_x$ model predicts a small different in reaction time as a function of PE (C) not seen in the empirical data, but the magnitude of the RT difference across PE levels is small relative to the magnitude of the RT difference across difficulty levels (higher d' leads to much faster RT). In contrast, while the alternative $C_\delta$ model can reproduce the differential confidence as a function of PE level given matched d' (D), it fails to capture the overlapping relationship between d' and meta-d' (E) and predicts overly large RT differences as a function of PE relative to the impact of d' on RT (F). Thus, the main $C_x$ model overall produces much closer fit to the data. Note that in the empirical data, there were only two levels of task difficulty; these correspond to the first and last data point in each linear curve in this plot. The second data point in each curve corresponds to the mean d' across difficulty levels in the Low PE condition, which was used for model fitting purposes (see Methods for details).

https://doi.org/10.1371/journal.pcbi.1008779.g004

$RT_{\text{low PE, difficult}}) \approx 30$ ms, which can be used as a reference point for what the empirical and modeled $\eta_{RT}$ effects translate to in terms of (RThigh PE, med—$RT_{\text{low PE, med}}$). In the empirical data, $\eta_{RT} = 0.044$ corresponds to an effect of PE level on RT of approximately 0.044*30 ms = 1.3 ms, whereas the equivalent RT effects in the model would be 11.2, 8.4, and 4.7 ms for $\alpha_{\text{low PE}} = 0.1$, 0.3, and 0.5 respectively. Thus, the magnitude of the incorrect prediction by the $C_x$ model for the effect of PE level on RT ranges from about 3.5–10 ms. While it is notable that the model seemingly makes a qualitatively incorrect prediction here, the magnitude of the effect is small enough that it does not pose a prohibitive failure. It is also possible that the apparent lack of an RT effect in the data is a false negative, which could plausibly occur if the true effect size were 10 ms or less.

By contrast, at levels of $\alpha_{\text{low PE}}$ and $\alpha_{\text{high PE}}$ that yield a reasonably close fit to the effect of PE level on confidence, the alternative $C_\delta$ model fares considerably worse in its predictions for meta-d' and RT. It incorrectly predicts that meta-d' should be higher in the High PE condition (PE effect on meta-d' = -0.029 in the data; 0.248, 0.181, and 0.114 in the model for $\alpha_{\text{low PE}}$ = 0.1, 0.3, and 0.5 respectively), and vastly overestimates the effect of PE level on RT ($\eta_{\text{RT}}$ = 0.044 in the data; 2.010, 1.539, 0.980 in the model for $\alpha_{\text{low PE}}$ = 0.1, 0.3, and 0.5 respectively). By the same logic described above, these model predictions correspond to predicted RT effects due to PE level in the $C_\delta$ model of 60.3, 46.2, and 29.4 ms for $\alpha_{\text{low PE}}$ = 0.1, 0.3, and 0.5 respectively. This means the magnitude of the incorrect prediction by the $C_\delta$ model for the effect of PE level on RT ranges from 28 to 59 ms—errors 6–10 times larger than those of the $C_x$ model. Such large effect sizes are also statistically incompatible with observing an effect close to zero in a reasonably powered sample (Koizumi et al's data set consisted of 480 trials). Side-by-side comparisons of the fitting errors for the $C_x$ and $C_\delta$ models at the values of $\alpha_{\text{high PE}}$ yielding the best fit the effect of PE level on mean confidence are shown in Fig 3D, 3H and 3L.

To further illustrate model behavior, in Fig 4 we show fits to the mean confidence, meta-d', and RT data for the full PE level (High / Low) x task difficulty (Difficult / Easy) design, separately for the $C_x$ and $C_\delta$ models. Displayed fits are derived from simulations using $\alpha_{\text{low PE}}$ = 0.1 (the only level of $\alpha_{\text{low PE}}$ for which the alternative $C_\delta$ model could capture the effect of PE level on confidence within the explored range of $\alpha_{\text{high PE}}$ levels) and $\alpha_{\text{high PE}}$ = 0.2548 ($C_x$ model) or $\alpha_{\text{high PE}}$ = 0.8928 ($C_\delta$ model), with these $\alpha_{\text{high PE}}$ values chosen so as to best fit the effect of PE level on mean confidence (corresponding to black circles in Fig 3). Echoing the more general results of Fig 3, the results of Fig 4 demonstrate how although both models can achieve good fits to the effect of PE level on performance-matched confidence, the $C_x$ model is considerably more accurate in its corresponding predictions for meta-d' and RT. As a reminder, these differences between the $C_x$ and $C_\delta$ models regarding the effect of PE level on meta-d' and RT stem from the differences between these models in fitted stimulus drive $S_{\text{high PE}}$ and $S_{\text{high NE}}$ (controlled by $\alpha_{\text{high PE}}$ via the equation $S_{\text{high NE}} = \alpha_{\text{high PE}} * S_{\text{high PE}}$) needed to account for the main d' and confidence effects.

In summary, these simulations demonstrate that across a comprehensive range of PE and NE levels, the $C_x$ model provides the best overall account of the data, achieving a fit to the effect of PE level on confidence that also yields an accurate fit to the effect of PE level on meta-d', and a reasonable fit to the effect of PE level on RT. Furthermore, these fits are accomplished with reasonably small differences in PE levels (i.e. with $\alpha_{\text{high PE}}$ reasonably close to $\alpha_{\text{low PE}}$). By contrast, although the $C_\delta$ model can also achieve a close fit to the effect of PE level on confidence, its corresponding fits to the effect of PE level on meta-d' and especially RT are poor, and furthermore this fit can only be achieved by positing an implausibly large difference in PE level.

## 3.3 Simulating data from Koizumi, Maniscalco, & Lau 2015 [10] Experiment 2B

In their Experiment 2B, Koizumi et al [10] used an alternative method to that used in Experiment 1A to achieve performance-matched differences in confidence. Stimuli consisted of random dot kinematograms with a fraction of the dots moving coherently to the left or right on each trial. Subjects had to indicate motion direction and then rate confidence on a 4-point scale. "Positive Evidence" (PE) was defined as the number of coherently moving dots, with this quantity depending on both the fraction of coherently moving dots and dot density; the remaining, randomly moving dots constituted sensory noise.

It is important to note a disanalogy between the structure of Experiments 1A and 2B of Koizumi et al. In Experiment 1A, stimuli consisted of the superposition of Positive Evidence (a high contrast grating tilting left or right), Negative Evidence (a lower contrast grating tilted in the opposite direction), and noise (randomly chosen fluctuations for pixel intensities). The PE component provided perceptual evidence for the correct perceptual choice, the NE component provided perceptual evidence for the incorrect perceptual choice, and noise systematically favored neither choice. By contrast, in Experiment 2B, stimuli consisted only of PE (coherent leftward or rightward motion) and noise (random motion). The analogue of NE in this task would be weaker coherent motion in the direction opposite to the PE motion, but no such opposite coherent motion was present in these stimuli. (To preempt a possible confusion, we note that Koizumi et al. used the term "NE" to refer to the randomly moving dots in Experiment 2B; here we have decided to abandon this usage in favor of reserving the term "NE" specifically for systematic perceptual evidence for the incorrect perceptual choice.)

The key experimental manipulation in Experiment 2B was the introduction of High PE and Low PE conditions; the fraction of coherently moving dots was equal across conditions, but dot density was twice as high in the High PE condition, thus yielding a higher absolute number of coherently moving dots. Additionally, there were two levels of task difficulty at each level of PE, thus giving rise to four experimental conditions in all: "High PE, Easy," High PE, Difficult," "Low PE, Easy," and Low PE, Difficult." Koizumi et al [10] found that the High PE and Low PE conditions yielded similar task performance (d'), but mean confidence was higher for High PE stimuli (Fig 4A of [10]).

As with the previous data sets, we predicted that the dissociation between d' and confidence in this experiment would be better captured by the main $C_x$ model than the alternative $C_\delta$ model, and tested this prediction by comparing model fits to the data. However, unlike the experimental results discussed above (Koizumi et al.'s [10] Experiment 1A), in this case the effect of performance-matched confidence was not driven by PE / NE manipulations; coherent motion always occurred in one direction only (PE), without any simultaneous coherent motion in the opposite direction (NE). Thus, it was necessary to use an alternative modeling strategy to the one used to model Koizumi et al.'s [10] Experiment 1A.

In the High PE condition of Experiment 2B, d' was similar to the Low PE condition due to having the same fraction of coherently moving dots, whereas confidence was higher due to having a higher absolute number of coherently moving dots. Thus, we reasoned that these data could be well modeled by supposing that the High PE condition has both stronger stimulus drive (more coherently moving dots) and stronger noise (more randomly moving dots); the increase in both stimulus strength and noise allows for signal-to-noise ratio (and hence d') to be similar, even as the same increases in signal and noise lead to higher absolute values of sensory evidence (and hence higher confidence). Indeed, previous work has demonstrated that increases in signal strength and noise can yield performance-matched increases in confidence under the assumption of fixed decision criteria [38], an assumption that is supported empirically in cases where the experimental conditions modeled with lower and higher variance in sensory evidence are randomly interleaved across trials [17]. We implemented this idea in the modeling by positing that the parameter controlling standard deviation of instantaneous sensory evidence, σ, would take on a larger value in the High PE condition.

To fit the $C_x$ and $C_\delta$ models to the data of Koizumi et al.'s [10] Experiment 2B, we adopted a similar model fitting approach as described above for their Experiment 1A, as detailed below.

To attain a comprehensive understanding of model behavior across different levels of sensory evidence variability, we fixed $\sigma_{\text{low PE}} = 0.1$ in the Low PE condition (the same value for σ used in all simulations described above) and explored model behavior when using values of

$\sigma_{\mathrm{high\ PE}}$ = 0.11, 0.12, . . ., 0.2 in the High PE condition (i.e. ranging from 10% to 100% increases in σ). For each pairing of $\sigma_{\mathrm{low\ PE}}$ and $\sigma_{\mathrm{high\ PE}}$, we fit the model parameters to the average d' and meta-d' values in the Low PE, Easy and Low PE, Difficult conditions and observed how this influenced the difference in mean confidence between the simulated High PE and Low PE conditions (Fig 5A).

Simulations results revealed that the main $C_x$ model can capture a wider range of differences in confidence due to manipulation of σ, which allows it to capture the observed magnitude of the (High PE confidence–Low PE confidence) effects using a relatively small increment in the corresponding $\sigma_{\mathrm{high\ PE}}$ (for $\sigma_{\mathrm{low\ PE}}$ = 0.1, the effect of PE level on confidence is captured by setting $\sigma_{\mathrm{high\ PE}}$ = 0.1087; Fig 5A). By contrast, the alternative $C_\delta$ model predicts only a comparatively small influence of σ on mean confidence, requiring a value of $\sigma_{\mathrm{high\ PE}}$ = 0.1679 to account for the confidence (Fig 5A). Thus, to achieve the same effect on performance-matched confidence, the $C_x$ model posits only a ~9% increase in sensory noise, whereas the $C_\delta$ model requires a much more substantial ~68% increase in noise.

To provide a more comprehensive characterization of the model fits to the empirical data, we also compared empirical and simulated data for the differences in reaction time and meta-d' induced by the High PE and Low PE conditions. RT and meta-d' data were not reported in Koizumi et al. [10]; here we re-analyzed their data set and report the RT and meta-d' data for the first time (Fig 6). In the empirical data, meta-d' increased and RT decreased with increasing d'. Additionally, PE level appeared to have a modest effect on both meta-d' and RT, with the high PE condition having slightly lower meta-d' and faster RT.

Simulation results for (High PE meta-d'–Low PE meta-d') and $\eta_{\mathrm{RT}}$ are shown in Fig 5B and 5C, respectively. As before, the point of main interest is how these models predict meta-d' and RT to behave at the specific values of $\sigma_{\mathrm{high\ PE}}$ that yield differences in mean confidence matching those observed in the data. These points are marked by filled circles and vertical dashed lines in Fig 5. For purposes of comparison to simulated data, we used linear fits to interpolate the empirical meta-d' and RT values at d' = 0.99 (corresponding to the average d' across the Low PE, Easy and Low PE, Difficult conditions) and computed the difference in the interpolated values for the High PE and Low PE conditions, yielding empirical values of High PE meta-d'–Low PE meta-d' = -0.129 and $\eta_{\mathrm{RT}}$ = 0.898 (horizontal dashed lines in Fig 5).

At the level of $\sigma_{\mathrm{high\ PE}}$ that yields a perfect fit to the effect of PE level on confidence (black circles in Fig 5A), the main $C_x$ model incorrectly predicts that meta-d' should be higher under High PE, although the magnitude of this incorrect prediction is modest (PE effect on meta-d' = -0.129 in the data; 0.095 in the model). Conversely, the $C_x$ model correctly predicts that RT should be faster under High PE than Low PE, and the predicted magnitude of this effect is comparable to the empirically observed effect ($\eta_{\mathrm{RT}}$ = 0.898 in the data; 1.252 in the model).

In the empirical data ($\mathrm{RT}_{\mathrm{low\ PE,\ easy}}$—$\mathrm{RT}_{\mathrm{low\ PE,\ difficult}}$) ≈ 10 ms, which can be used as a reference point for what the empirical and modeled $\eta_{\mathrm{RT}}$ effects translate to in terms of ($\mathrm{RT}_{\mathrm{high\ PE,\ med}}$—$\mathrm{RT}_{\mathrm{low\ PE,\ med}}$). In the empirical data, $\eta_{\mathrm{RT}}$ = 0.898 corresponds to an effect of PE on RT of approximately 0.898*10 ms = 8.98 ms, whereas the equivalent RT effect in the model would be 12.52 ms. Thus, the error in the the $C_x$ model's prediction for the effect of PE level on RT amounts to about 3.5 ms.

By contrast, at the level of $\sigma_{\mathrm{high\ PE}}$ that yield a perfect fit to the effect of PE level on confidence, the alternative $C_\delta$ model fares considerably worse in its predictions for both meta-d' and RT. Like the $C_x$ model, the $C_\delta$ model incorrectly predicts that meta-d' should be higher in the High PE condition, but the magnitude of this incorrect prediction is 5 times as large as in the $C_x$ model (PE effect on meta-d' = -0.129 in the data; 0.518 in the model); indeed, the predicted meta-d' in the High PE condition is considerably higher than d', which violates a theoretically expected, and commonly empirically observed, constraint such that meta-d' ≤ d' [29].

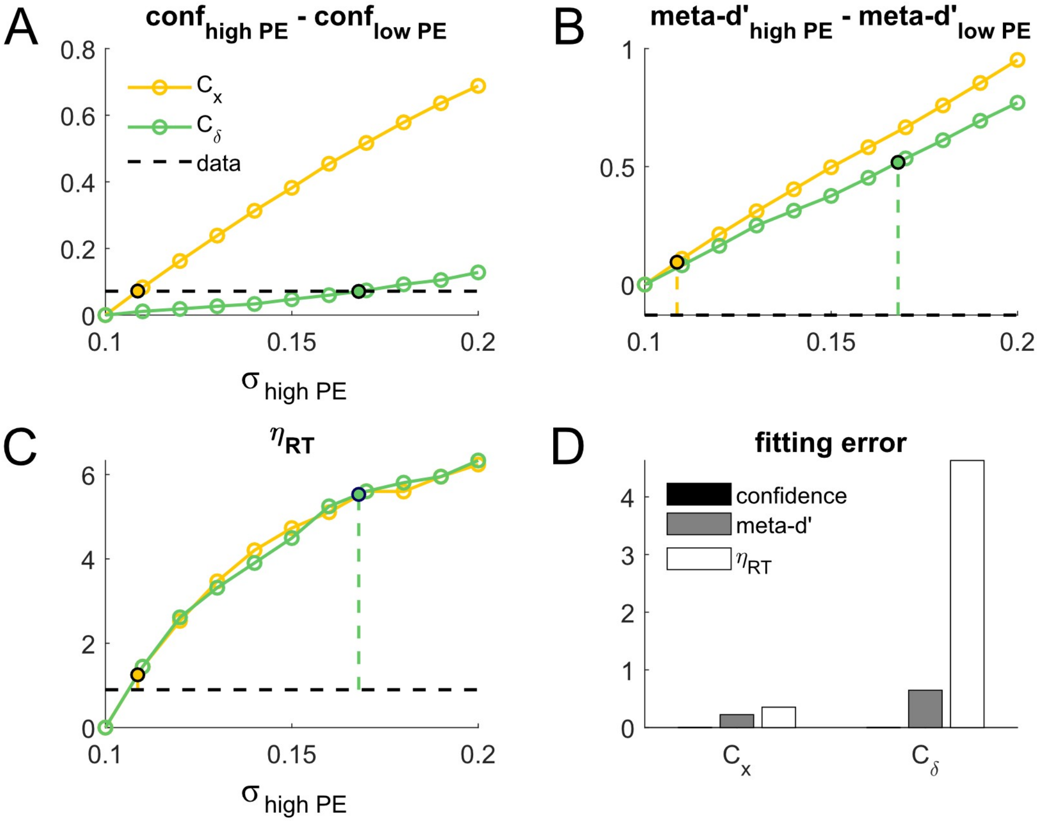

**Fig 5. Comprehensive simulations showing fitting procedures for the main $C_x$ model and alternative $C_\delta$ model for Koizumi et al.'s [10] Experiment 2B.** As in the analysis of Koizumi et al.'s Experiment 1A in Fig 3, we examined performance-matched confidence differences, performance-matched meta-d' differences, and relative RT effect sizes as a function of PE level. (A) For each of the $C_x$ and $C_\delta$ models, we determined the level of sensory evidence noise in the High PE ($\sigma_{\text{high PE}}$) condition that best matched the observed magnitude of confidence differences in the empirical data (dashed horizontal line), marked with black circles. We then "read out" the predicted meta-d' differences (B) and relative RT effect sizes (C). (D) We also computed fitting error for each model as the difference between the simulated and empirical confidence differences, meta-d' differences, and relative RT effects (D). As with Experiment 1A, the $C_x$ model produced a good fit to the empirical data across the entire spectrum of measures considered, whereas the alternative $C_\delta$ model produced substantial errors for predicted meta-d' and RT when made to match the empirically-observed confidence differences.

https://doi.org/10.1371/journal.pcbi.1008779.g005

Similarly, the $C_\delta$ model vastly overestimates the effect of PE level on RT ($\eta_{RT}$ = 0.898 in the data; 5.530 in the model). By the same logic described above, this $\eta_{RT}$ value corresponds to a predicted effect of PE level on RT of 55.3 ms. Thus, the magnitude of the incorrect prediction by the $C_\delta$ model for the effect of PE level on RT is about 45 ms—an error over 10 times larger than that of the $C_x$ model (3.5 ms). Side-by-side comparisons of the fitting errors for the $C_x$ and $C_\delta$ models at the values of $\sigma_{\text{high PE}}$ yielding the best fit the effect of PE level on mean confidence are shown in Fig 5D.

To further illustrate model behavior, in Fig 6 we show fits to the mean confidence, meta-d', and RT data for the full PE level (High / Low) x task difficulty (Difficult / Easy) design, separately for the $C_x$ and $C_\delta$ models. Displayed fits are derived from simulations using $\sigma_{\text{low PE}}$ = 0.1 and $\sigma_{\text{high PE}}$ = 0.1087 ($C_x$ model) or $\sigma_{\text{high PE}}$ = 0.1679 ($C_\delta$ model), with these $\sigma_{\text{high PE}}$ values chosen so as to best fit the effect of PE level on mean confidence (corresponding to black circles in Fig 5). Echoing the more general results of Fig 5, the results of Fig 6 demonstrate how although both models can achieve good fits to the effect of PE level on performance-matched

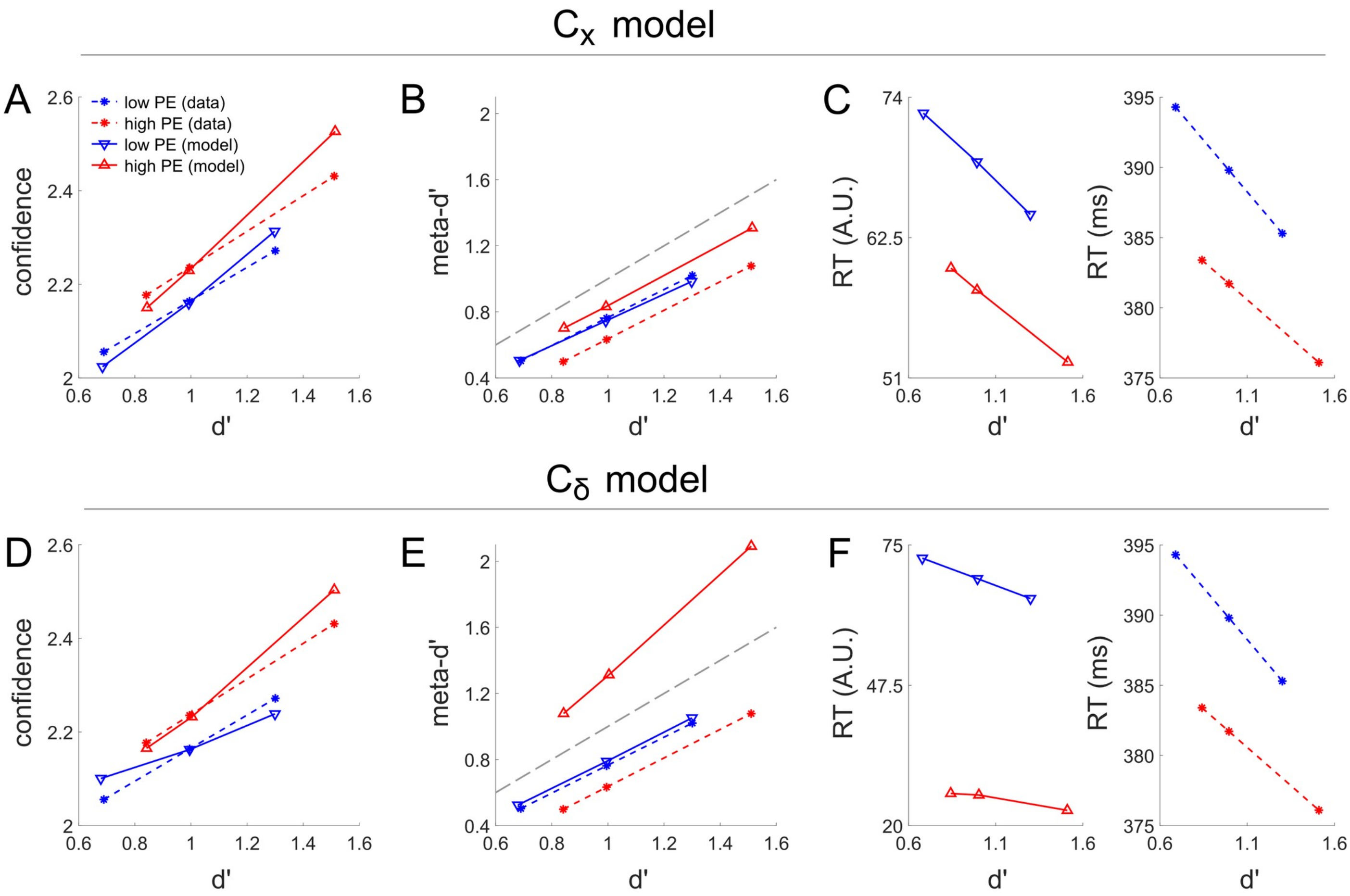

**Fig 6. Modeling results for the effects reported by Koizumi and colleagues [10] in their Experiment 2B with parameter values chosen to best capture the effect of PE level on mean confidence (corresponding to the data points shown in black circles in Fig 5).** As for their Experiment 1A, the fitted main $C_x$ model reproduces the differential confidence for matched d' performance for high versus low PE stimuli (A) and the relatively overlapping distributions of meta-d' as a function of d' (B), although somewhat less well than for experiment 1. However, in contrast to Experiment 1A the main $C_x$ model now correctly predicts the relationship between RT differences a s function of d' and PE (C) much better than for Experiment 1A. In contrast, while the alternative $C_\delta$ model can reproduce the differential confidence as a function of density level given matched d' (D), it again badly fails to capture the overlapping relationship between d' and meta-d' (E)–this time predicting an implausibly high meta-d' value. The alternative $C_\delta$ model also again predicts overly large RT differences as a function of density level, while predicting almost zero impact of d' on RT (F). Thus, as with Experiment 1A, the main $C_x$ model overall produces much closer fit to the data. Note that in the empirical data, there were only two levels of task difficulty; these correspond to the first and last data point in each linear curve in this plot. The second data point in each curve corresponds to the mean d' across difficulty levels in the Low PE condition, which was used for model fitting purposes (see Methods for details).

https://doi.org/10.1371/journal.pcbi.1008779.g006

confidence, the main $C_x$ model is considerably more accurate in its corresponding predictions for meta-d' and RT. Here, as above, differences in meta-d' and RT between the $C_x$ and $C_\delta$ models result from differences in fitted stimulus drive $S_{PE}$ for each model, which was needed to account for the main d' and confidence effects.

In summary, these simulations demonstrate that across a comprehensive range of σ levels, the main $C_x$ model provides the best overall account of the data, achieving a fit to the effect of PE level on confidence that also yields a reasonably close fit to the effect of PE level on meta-d' (albeit one that goes in the wrong direction), as well as a close fit to the effect of PE level on RT. Furthermore, these fits are accomplished with a small (~9%), and therefore plausible, increase in σ. By contrast, although the alternative $C_\delta$ model can also achieve a close fit to the effect of PE level on confidence, its corresponding fits to the effect of PE level on meta-d' and

RT are both very poor, and its prediction for meta-d' in the High PE condition is implausibly high. Furthermore, the fit to the confidence effect can only be achieved by positing a relatively large increase (~68%) in σ.

## 4. Discussion

How the brain calculates subjective decision confidence is still a topic of active debate [1,2,39–43]. Although dominant models suggest that confidence reflects an optimal readout of the probability that a decision is correct [1–9], it appears challenging for such models to account for counterintuitive behaviors in which confidence and accuracy dissociate [10,11,13,14,16,17,44]. An alternative hypothesis suggesting that confidence reflects a heuristic reliance on decision-congruent evidence [10,14,16,18,21] captures many of these behaviors, and is supported by human intracranial electrophysiology [23].

Here, we considered how decision-congruent computations of perceptual confidence might be biologically implemented based on known properties of perceptual circuitry. We hypothesized that tuned inhibition (a generalization of the concept of tuned normalization) [24–26] differentially influences a neuron's role in perceptual decision-making and confidence, such that more inhibited units (corresponding to the net evidence for a perceptual choice) drive decisions and less inhibited units (corresponding to decision-congruent evidence) drive confidence. We developed the Differential Tuned Inhibition model to test this hypothesis. Our results show that such a network can explain counterintuitive behaviors reported in the literature [4,10,12–14,16–18,44,45]. We further demonstrate that the model's special property of weighting less inhibited units more heavily in computing confidence is the key to capturing empirical findings, since control simulations demonstrate that the model fails to reproduce these findings when instead *more* inhibited units drive confidence. This provides preliminary but converging evidence that decision-congruent confidence computations may be implemented via tuned inhibition.

It might be argued that some over-simplified optimal diffusion-type models [6,7,9] should not be expected to account for counterintuitive behaviors due to their simplicity. A recent modification of these optimal diffusion-type models suggests that the optimal perceptual confidence readout must also depend on the time it took for evidence to accumulate [3–5]. Although it has been suggested that neurons in lateral intraparietal cortex (LIP) may encode elapsed time [46–48], these neurons' activity has not yet been causally or directly linked to subjective confidence (see also [49]). This suggests that how this time-dependent diffusion framework might be biologically implemented is nontrivial, inspiring the work presented here.

We employed a two-stage evidence accumulation model [9] in which, after the initial perceptual decision is triggered by accumulated evidence surpassing the decision threshold, accumulation continues for τ additional time steps. As τ increases, confidence ratings become more diagnostic of task accuracy, and thus meta-d' increases (S1C and S1D Fig). This model structure thus has the considerable benefit of naturally lending itself to fitting meta-d', which (in our experience) sometimes proves a difficult task for dynamic evidence accumulation models.

We adopted a simple but powerful approach for model fitting and model comparison. After fixing sensory noise σ and decision threshold T to *a priori* values, stimulus drive S could be chosen to exactly match all empirical d' values, and (given the fitted value of τ), confidence thresholds $U_r$ could be chosen to exactly match overall empirical distributions of confidence ratings (i.e. P(conf = i) where $1 \leq i \leq 4$ for a 4-point rating scale). Against this backdrop (which already perfectly captures much of the data), our general strategy was to fit τ to the meta-d' value at a single data point and, using this fitted value of τ in all other conditions, observe what patterns emerged in the confidence data across all other data points as a consequence of the structure of the simulation (i.e., the differing stimulus strengths for the "B" stimulus in the simulations of Maniscalco et al [14]; the differing levels of positive and negative evidence in the simulations of Koizumi et al [10] Experiment 1A; the differing levels of sensory evidence noise in the simulations of Koizumi et al [10] Experiment 2B). Our key emphasis was thus not to derive exact fits for every data point, but rather to assess qualitative patterns in the fits that emerged from a simple modeling approach and compare how the fits differed for the main Differential Tuned Inhibition model (i.e. the $C_x$ model) and the alternative $C_\delta$ model.

The counterintuitive empirical findings of Maniscalco et al [14], which show that in some conditions metacognitive sensitivity can decrease in spite of increasing task performance, provide a strong test for models of perceptual confidence, since such models usually naturally predict that metacognitive sensitivity and task performance positively correlate (and indeed, such positive correlations are widely observed in empirical data). We found that simulations of this data set unambiguously favored the main $C_x$ model over the alternative $C_\delta$ model, as only the former was capable of reproducing the distinctive X-shaped dissociation in response-conditional meta-d' curves, whereby meta-d' for "S1" responses decreases and meta-d' for "S2" responses increases with increases in d' (Fig 2).

In the same paper as modeled here, Maniscalco and colleagues [14] also examined the effects of feedback on metacognitive sensitivity (meta-d'). They found that feedback on task accuracy and confidence judgments led to the near-disappearance of the X-shaped crossover effect in response-conditional meta-d', instead producing results resembling those of the $C_\delta$ model presented here (our Fig 2B, their Fig 7). These observations suggest that decision-congruent evidence biases in perceptual confidence can change fluidly depending on training or other factors, suggesting promising avenues for future studies testing the Differential Tuned Inhibition model with paradigms designed to manipulate response-conditional meta-d'. We note that our model formulation does not explicitly specify whether the confidence readout mechanism may be innate or learned through the lifetime, and therefore is not committed to viewing the readout as inflexible or hardwired versus the possibility of accommodating changes due to learning or other factors. Future research could investigate an expanded version of the model in which the confidence readout mechanism, and its potential change as a function of learning, is explicitly modeled, as well as exploring whether it may rest on innate versus learned mechanisms.

At first glance, the simulation results for Koizumi et al [10] Experiments 1A and 2B were more equivocal, since both the $C_x$ and $C_\delta$ models could capture the dissociations in these data whereby some conditions exhibit different levels of mean confidence in spite of having similar task performance (Figs 4A, 4D, 6A and 6D). However, the alternative $C_\delta$ model required implausible stimulus manipulations many times stronger than those required by the main $C_x$ model to capture the performance-matched confidence effects. Furthermore, the parameter settings that allowed the $C_\delta$ model to achieve this fit also entailed predicted effects of reaction time and meta-d' that were strongly incorrect, with error magnitudes many times as large as the corresponding predictions for the $C_x$ model (Figs 3–6), including an implausibly large predicted value for meta-d' such that meta-d' > d' (Fig 6E). By contrast, the $C_x$ model fits to meta-d' and RT, while not perfect, were all reasonably close to the empirically observed data—a finding made more impressive by the fact that the model structure and parameter values were not chosen in any way so as to provide reasonable fits to meta-d' and RT, but rather these fits naturally "fell out" of the simulation structure intended to match salient features of experimental design and parameter values intended to fit the d' and mean confidence data. Thus, on balance, the results of the Koizumi et al [10] simulations strongly favored the main $C_x$ model.

Thus, we found that by considering the models' ability to capture a comprehensive set of behavioral data—including d', RT, mean confidence, and meta-d'—we were better able to distinguish the effectiveness of competing models in capturing those data. We especially wish to highlight the utility and power of metacognitive sensitivity (as measured here by meta-d' [29]) as a target for model fitting. In addition to capturing task performance, reaction time, or mean confidence within an aggregate of trials, any successful model of perceptual confidence should also be able to account for metacognitive sensitivity—i.e., the trial-by-trial correspondence between confidence and accuracy. As we have shown here, assessing fit to metacognitive sensitivity data can be an incisive tool for model evaluation and comparison—yet the vast majority of extant studies on dynamic decision making models of perceptual confidence do not consider patterns of metacognitive sensitivity in the data or how the model can (or cannot) account for them. The simulations of the Koizumi et al [10] experiments in the present study are instructive in that the competing models appeared equal in their ability to account for performance-matched confidence, but yet could be distinguished by their relative ability to account for meta-d'.

Notably, although we only performed model simulations for three sets of experimental results, the simulation findings apply more broadly to any experiments using similar manipulations. Several other studies have used manipulations of positive and negative evidence to achieve performance-matched confidence of the sort employed by Koizumi et al [10] Experiment 1A [15,16,45]. The simulations of Koizumi et al [10] Experiment 2B are perhaps even more broadly informative, insofar as they may shed light on any experimental design that can be modeled as influencing task performance and confidence by means of altering the variability of sensory evidence—including studies employing stimulus manipulations [5] but also manipulations of attention [13] and direct intervention on neural activity [11,12]. A recent model with similar flavor to ours proposed a competing accumulator framework in which both confidence and decision were read out from the same units, but these units were only partially inhibited [21]; this model captured effects wherein confidence appeared to rely more on decision-congruent evidence, but decisions on a balance of evidence between decision-congruent and decision-incongruent evidence. However, in that paper the authors did not explore whether such a model could capture meta-d'—and as we have shown here the summary behaviors of d' and confidence can be explained by an alternative model in which the same units drive decisions and confidence but, critically, meta-d' could not be captured by such a model. Future work should explore the degree to which strength of feed-forward inhibition might produce more nuanced behaviors, as our goal with the current model was to provide a proof of concept that tuned inhibition in sensory circuits can provide a biologically plausible mechanism for implementing decision-congruent confidence computations.

Our results suggest a potential adaptive consequence for the presence of tuned inhibition [24–26] within a behaving organism: the presence of both more and less inhibited neurons within a perceptual decision-making circuit may allow an organism to better solve both fine-grained discrimination and detection tasks. When making fine-grained discrimination or identification judgments about an object or stimulus ("What is that thing?"), a useful strategy would be to rely on a system that is not as sensitive to random fluctuations, i.e. a more strongly inhibited system. But when making detection decisions ("Is there something out there?"), such strong inhibition would be highly undesirable, so a useful strategy would be to rely on less inhibited parts of the network. Both of these tasks are important for an organism to execute, and so it seems beneficial that a system might retain elements that can preferentially contribute to each task rather than implementing only one.

The question then becomes why the system would recruit the 'detection' portions of its circuitry to compute confidence, specifically relying on the magnitude of decision-congruent evidence. One reason may simply be heuristic, that the detectability and identifiability of a stimulus are often correlated in the real world; although in laboratory conditions these can certainly be dissociated, in real-world conditions they often go hand in hand. Indeed, it has been noted that the width of the posterior distribution in a probabilistic population code [42,50] covaries with the overall firing rate of a population [51]; less inhibited 'detection' neurons would more readily affect a population's overall firing rate, suggesting a potential neural substrate for this heuristic. Perhaps due to this strong statistical coincidence in naturalistic environments, the system did not need to evolve away from such a heuristic, which also conveniently minimizes the need to retain information about unchosen stimulus identity possibilities once a perceptual inference has been made [23] and therefore might be computationally efficient. Indeed, such over-reliance on decision-congruent evidence—i.e., a "confirmation bias" [52]—has also been observed in other post-decisional (non-metacognitive) perceptual judgments [19,20], value judgments [53–58], and metamemory [59,60], suggesting it may be a domain-general strategy that serves also to reduce cognitive dissonance and improve self-consistency.

Using the absolute strength of decision-congruent evidence to judge confidence could also indicate that a confidence judgment attempts to infer the possible *cause* of the signals that led to the perceptual inference as externally- or internally-generated [61–63]: Are these signals strong enough to indicate an external stimulus, or are they likely to simply reflect internal noise fluctuations? A mechanism that keeps track of the absolute amount of evidence, regardless of the balance, would be critical to successfully solving such a causal inference problem by allowing the system to differentiate between strong versus weak signals even when the signals themselves are equally ambiguous (i.e., equally favor one versus another possible stimulus identity). And finally, that 'detection' circuitry might contribute to metacognitive judgments is also supported by reports of neurons coding for the detectability (or lack thereof) of a stimulus in prefrontal cortex [64], an area known to be involved in metacognitive computations (including judgments of 'visibility, i.e. awareness) in perception and memory [39,65–75].

Because the above results are suggestive in nature, confirmation that tuned inhibition is utilized in perceptual decisions versus confidence judgments as hypothesized here will critically depend on experiments designed to reveal the biological mechanism in awake, behaving animals. As an initial exploratory test, however, we capitalized on an existing dataset consisting of electrophysiological recordings in Rhesus macaque superior colliculus, a subcortical area involved in perceptual decision-making [76–79] and containing the type of evidence accumulation neurons typically assumed to be involved in perceptual decision-making [3,6,7,9,35,40,80] (S4 Text); previously, tuned inhibition (tuned normalization) has only been reported in areas not involved in evidence accumulation [24–26]. Using catch trials from those data, we see preliminary support for one of our model's critical predictions: that evidence accumulation neurons in perceptual decision-making areas ought also to exhibit tuned inhibition. Additional detail is provided in S4 Text and S8 Fig. Future work should extend these preliminary analyses and design specific experiments to test and arbitrate among our model and competing models of perceptual confidence [3–9].

Here we demonstrated that inhibition tuning provides a biologically plausible mechanism for implementing confidence computations that demonstrate an over-reliance on decision-congruent information. Our findings lead to testable hypotheses about the role of tuned inhibition in a neuron's contribution to a decision versus a confidence judgment: activity of more inhibited units should reflect an observer's objective decisions more than confidence judgments, while the opposite should be true for less inhibited neurons. Future electrophysiological studies should further explore the extent to which this hypothesis can be verified in the neurobiology of perceptual decision-making circuitry. It has also been reported that tuned inhibition is spatially ‘clumped’, i.e. that nearby neurons have more similar inhibition profiles than neurons separated by longer distances [25]. The present findings thus pave the way for noninvasive neuroscience techniques, such as spatially coarser functional MRI in humans, to clarify the role of inhibition tuning in perceptual and cognitive decisions and metacognitive evaluations of these choices.

## Supporting information

**S1 Text. Full methods for model fitting.**
(DOCX)

**S2 Text. Additional simulation results and comparison of the $C_x$ model to the 2D-SDT model of Maniscalco, Peters, & Lau 2016.**
(DOCX)

**S3 Text. Testing a Leaky Competing Accumulator implementation of the model on the data of Maniscalco, Peters, & Lau 2016.**
(DOCX)

**S4 Text. Evidence for tuned inhibition in macaque superior colliculus.**
(DOCX)

**S1 Fig. Parameter fitting for the simulations of Maniscalco, Peters, & Lau 2016.**
(TIF)

**S2 Fig. Data and simulations for mean confidence as a function of S2 stimulus strength, perceptual decision, and accuracy.**
(TIF)

**S3 Fig. Simulated distributions of evidence in confidence units at the time of confidence rating as a function of stimulus and accuracy.**
(TIF)

**S4 Fig. LCA model structure.**
(TIF)

**S5 Fig. Simulation results for the model shown in S4A Fig.**
(TIF)

**S6 Fig. Simulated distributions of evidence in confidence units at the time of confidence rating as a function of stimulus and accuracy for the LCA model shown in S4A Fig.**
(TIF)

**S7 Fig. Simulation results for the model shown in S4B Fig.**
(TIF)

**S8 Fig. Evidence for tuned inhibition in macaque superior colliculus (SC).**
(TIF)

## Acknowledgments

We thank Bosco Tjan for seminal discussion on the theoretical foundations for this project.

## Author Contributions

**Conceptualization:** Brian Maniscalco, Hakwan Lau, Megan A. K. Peters.

**Data curation:** Brian Maniscalco, Megan A. K. Peters.

**Formal analysis:** Brian Maniscalco, Brian Odegaard, Piercesare Grimaldi, Seong Hah Cho, Megan A. K. Peters.

**Funding acquisition:** Michele A. Basso, Hakwan Lau, Megan A. K. Peters.

**Investigation:** Piercesare Grimaldi, Michele A. Basso.

**Methodology:** Brian Maniscalco, Hakwan Lau, Megan A. K. Peters.

**Project administration:** Megan A. K. Peters.

**Resources:** Michele A. Basso.

**Software:** Brian Maniscalco, Megan A. K. Peters.

**Supervision:** Michele A. Basso, Hakwan Lau, Megan A. K. Peters.

**Visualization:** Brian Maniscalco, Brian Odegaard, Piercesare Grimaldi, Seong Hah Cho, Megan A. K. Peters.

**Writing – original draft:** Brian Maniscalco, Hakwan Lau, Megan A. K. Peters.

**Writing – review & editing:** Brian Maniscalco, Hakwan Lau, Megan A. K. Peters.

## References

1. Sanders JI, Hangya B, Kepecs A. Signatures of a Statistical Computation in the Human Sense of Confidence. Neuron. 2016; 90: 499–506. https://doi.org/10.1016/j.neuron.2016.03.025 PMID: 27151640
2. Pouget A, Drugowitsch J, Kepecs A. Confidence and certainty: distinct probabilistic quantities for different goals. Nat Neurosci. 2016; 19: 366–374. https://doi.org/10.1038/nn.4240 PMID: 26906503
3. Kiani R, Corthell L, Shadlen MN. Choice Certainty Is Informed by Both Evidence and Decision Time. Neuron. 2014; 84: 1329–1342. https://doi.org/10.1016/j.neuron.2014.12.015 PMID: 25521381
4. Fetsch CR, Kiani R, Newsome WT, Shadlen MN. Effects of Cortical Microstimulation on Confidence in a Perceptual Decision. Neuron. 2014; 1–8.
5. Zylberberg A, Fetsch CR, Shadlen MN, Frank MJ. The influence of evidence volatility on choice, reaction time and confidence in a perceptual decision. eLife. 2016; 5: e17688. https://doi.org/10.7554/eLife.17688 PMID: 27787198
6. Ratcliff R, McKoon G. The diffusion decision model: Theory and data for two-choice decision tasks. Neural Comput. 2008; 20: 873–922. https://doi.org/10.1162/neco.2008.12-06-420 PMID: 18085991
7. Ratcliff R, Rouder JN. Modeling Response Times for Two-Choice Decisions. Psychol Sci. 1998; 9: 347–356.
8. Tsetsos K, Gao J, McClelland JL, Usher M. Using Time-Varying Evidence to Test Models of Decision Dynamics: Bounded Diffusion vs. the Leaky Competing Accumulator Model. Front Neurosci. 2012; 6. https://doi.org/10.3389/fnins.2012.00079 PMID: 22701399
9. Pleskac TJ, Busemeyer JR. Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. Psychol Rev. 2010; 117: 864–901. https://doi.org/10.1037/a0019737 PMID: 20658856
10. Koizumi A, Maniscalco B, Lau H. Does perceptual confidence facilitate cognitive control? Atten Percept Psychophys. 2015. https://doi.org/10.3758/s13414-015-0843-3 PMID: 25737256
11. Rahnev D, Maniscalco B, Luber B, Lau H, Lisanby SH. Direct injection of noise to the visual cortex decreases accuracy but increases decision confidence. J Neurophysiol. 2012; 107: 1556–1563. https://doi.org/10.1152/jn.00985.2011 PMID: 22170965
12. Peters MAK, Fesi J, Amendi N, Knotts JD, Lau H, Ro T. Transcranial magnetic stimulation to visual cortex induces suboptimal introspection. Cortex. 2017; 93: 119–132. https://doi.org/10.1016/j.cortex.2017.05.017 PMID: 28646672
13. Rahnev D, Maniscalco B, Graves T, Huang E, de Lange FP, Lau H. Attention induces conservative subjective biases in visual perception. Nat Neurosci. 2011; 14: 1513–1515. https://doi.org/10.1038/nn.2948 PMID: 22019729

**14.** Maniscalco B, Peters MAK, Lau H. Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity. Atten Percept Psychophys. 2016. https://doi.org/10.3758/s13414-016-1059-x PMID: 26791233

**15.** Odegaard B, Grimaldi P, Cho SH, Peters MAK, Lau H, Basso MA. Superior colliculus neuronal ensemble activity signals optimal rather than subjective confidence. Proceedings of the National Academy of Sciences. 2018; 115: E1588–E1597. https://doi.org/10.1073/pnas.1711628115 PMID: 29382765

**16.** Samaha J, Barrett JJ, Sheldon AD, LaRocque JJ, Postle BR. Dissociating Perceptual Confidence from Discrimination Accuracy Reveals No Influence of Metacognitive Awareness on Working Memory. Front Psychol. 2016; 7: 851. https://doi.org/10.3389/fpsyg.2016.00851 PMID: 27375529

**17.** Maniscalco B, Castaneda OG, Odegaard B, Morales J, Rajananda S, Peters MAK. The metaperceptual function: Exploring dissociations between confidence and task performance with type 2 psychometric curves. 2020. https://doi.org/10.31234/osf.io/5qrjn

**18.** Samaha J, Iemi L, Postle BR. Prestimulus alpha-band power biases visual discrimination confidence, but not accuracy. Conscious Cogn. 2017. https://doi.org/10.1016/j.concog.2017.02.005 PMID: 28222937

**19.** Stocker AA, Simoncelli EP. A Bayesian Model of Conditioned Perception. Adv Neural Inf Process Syst. 2008; 20: 1409–1416.

**20.** Luu L, Stocker AA. Post-decision biases reveal a self-consistency principle in perceptual inference. eLife Sciences. 2018; 7: e33334. https://doi.org/10.7554/eLife.33334 PMID: 29785928

**21.** Zylberberg A, Barttfeld P, Sigman M. The construction of confidence in a perceptual decision. Front Integr Neurosci. 2012; 6: 79–79. https://doi.org/10.3389/fnint.2012.00079 PMID: 23049504

**22.** Aitchison L, Bang D, Bahrami B, Latham PE. Doubly Bayesian analysis of confidence in perceptual decision-making. PLoS Comput Biol. 2015; 11: e1004519–e1004519. https://doi.org/10.1371/journal.pcbi.1004519 PMID: 26517475

**23.** Peters MAK, Thesen T, Ko YD, Maniscalco B, Carlson C, Davidson M, et al. Perceptual confidence neglects decision-incongruent evidence in the brain. Nature Human Behaviour. 2017.

**24.** Ni AM, Ray S, Maunsell JHR. Tuned normalization explains the size of attention modulations. Neuron. 2012; 73: 803–813. https://doi.org/10.1016/j.neuron.2012.01.006 PMID: 22365552

**25.** Ruff DA, Alberts JJ, Cohen MR. Relating normalization to neuronal populations across cortical areas. J Neurophysiol. 2016; 4: jn.00017.2016-jn.00017.2016. https://doi.org/10.1152/jn.00017.2016 PMID: 27358313

**26.** Verhoef B-E, Maunsell JHR. Attention-related changes in correlated neuronal activity arise from normalization mechanisms. Nat Neurosci. 2017. https://doi.org/10.1038/nn.4572 PMID: 28553943

**27.** Carandini M, Heeger DJ. Normalization as a canonical neural computation. Nat Rev Neurosci. 2012; 13: 51–62.

**28.** Reynolds JH, Heeger DJ. The normalization model of attention. Neuron. 2009; 61: 168–185. https://doi.org/10.1016/j.neuron.2009.01.002 PMID: 19186161

**29.** Maniscalco B, Lau H. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. Conscious Cogn. 2012; 21: 422–430. https://doi.org/10.1016/j.concog.2011.09.021 PMID: 22071269

**30.** Ohshiro T, Angelaki DE, DeAngelis GC. A normalization model of multisensory integration. Nat Neurosci. 2011; 14: 775–782. https://doi.org/10.1038/nn.2815 PMID: 21552274

**31.** Churchland AK. Normalizing relations between the senses. Nat Neurosci. 2011; 14: 672–673. https://doi.org/10.1038/nn.2850 PMID: 21613993

**32.** Ling S, Blake R. Normalization regulates competition for visual awareness. Neuron. 2012; 75: 531–540. https://doi.org/10.1016/j.neuron.2012.05.032 PMID: 22884335

**33.** Nassi JJ, Gómez-Laberge C, Kreiman G, Born RT. Corticocortical feedback increases the spatial extent of normalization. Front Syst Neurosci. 2014; 8: 105–105. https://doi.org/10.3389/fnsys.2014.00105 PMID: 24910596

**34.** Simoncelli EP, Heeger DJ. A model of neuronal responses in visual area MT. Vision Res. 1998; 38: 743–761. https://doi.org/10.1016/s0042-6989(97)00183-1 PMID: 9604103

**35.** Usher M, McClelland JL. The time course of perceptual choice: the leaky, competing accumulator model. Psychol Rev. 2001; 108: 550–592. https://doi.org/10.1037/0033-295x.108.3.550 PMID: 11488378

**36.** Yu S, Pleskac TJ, Zeigenfuse MD. Dynamics of Postdecisional Processing of Confidence. J Exp Psychol Gen. 2015; 144: 489–510. https://doi.org/10.1037/xge0000062 PMID: 25844627

37. Ratcliff R, Rouder JN. A diffusion model account of masking in two-choice letter identification. J Exp Psychol Hum Percept Perform. 2000; 26: 127–140. https://doi.org/10.1037//0096-1523.26.1.127 PMID: 10696609

38. Morales J, Odegaard B, Maniscalco B. The Neural Substrates of Conscious Perception without Performance Confounds. philpapers.org. 2020. Available: https://philpapers.org/rec/MORTNS-4

39. Fleming SM, Dolan RJ. The neural basis of metacognitive ability. Philos Trans R Soc Lond B Biol Sci. 2012; 367: 1338–1349. https://doi.org/10.1098/rstb.2011.0417 PMID: 22492751

40. Yeung N, Summerfield C. Metacognition in human decision-making: confidence and error monitoring. Philos Trans R Soc Lond B Biol Sci. 2012; 367: 1310–1321. https://doi.org/10.1098/rstb.2011.0416 PMID: 22492749

41. Charles L, Van Opstal F, Marti S, Dehaene S. Distinct brain mechanisms for conscious versus subliminal error detection. Neuroimage. 2013; 73: 80–94. https://doi.org/10.1016/j.neuroimage.2013.01.054 PMID: 23380166

42. Ma WJ, Beck JM, Latham P, Pouget A. Bayesian inference with probabilistic population codes. Nat Neurosci. 2006; 9: 1432–1438. https://doi.org/10.1038/nn1790 PMID: 17057707

43. Orbán G, Berkes P, Fiser J, Lengyel M. Neural Variability and Sampling-Based Probabilistic Representations in the Visual Cortex. Neuron. 2016; 92: 530–543. https://doi.org/10.1016/j.neuron.2016.09.038 PMID: 27764674

44. Rahnev D, Bahdo L, de Lange FP, Lau H. Prestimulus hemodynamic activity in dorsal attention network is negatively associated with decision confidence in visual perception. J Neurophysiol. 2012; 108: 1529–1536. https://doi.org/10.1152/jn.00184.2012 PMID: 22723670

45. Samaha J, Switzky M, Postle BR. Confidence boosts serial dependence in orientation estimation. J Vis. 2019; 590. https://doi.org/10.1167/19.4.25 PMID: 31009526

46. Janssen P, Shadlen MN. A representation of the hazard rate of elapsed time in macaque area LIP. Nat Neurosci. 2005; 8: 234–241. https://doi.org/10.1038/nn1386 PMID: 15657597

47. Leon MI, Shadlen MN. Representation of time by neurons in the posterior parietal cortex of the macaque. Neuron. 2003; 38: 317–327. https://doi.org/10.1016/s0896-6273(03)00185-5 PMID: 12718864

48. Finnerty GT, Shadlen MN, Jazayeri M, Nobre AC, Buonomano DV. Time in Cortical Circuits. J Neurosci. 2015; 35: 13912–13916. https://doi.org/10.1523/JNEUROSCI.2654-15.2015 PMID: 26468192

49. Bang D, Fleming SM. Distinct encoding of decision confidence in human medial prefrontal cortex. Proc Natl Acad Sci U S A. 2018; 115: 6082–6087. https://doi.org/10.1073/pnas.1800795115 PMID: 29784814

50. Ma WJ, Beck JM, Pouget A. Spiking networks for Bayesian inference and choice. Curr Opin Neurobiol. 2008; 18: 217–222. https://doi.org/10.1016/j.conb.2008.07.004 PMID: 18678253

51. Bays PM. A signature of neural coding at human perceptual limits. J Vis. 2016; 16: 4. https://doi.org/10.1167/16.11.4 PMID: 27604067

52. Abrahamyan A, Silva LL, Dakin SC, Carandini M, Gardner JL. Adaptable history biases in human perceptual decisions. Proc Natl Acad Sci U S A. 2016; 113: E3548–57. https://doi.org/10.1073/pnas.1518786113 PMID: 27330086

53. Koster R, Duzel E, Dolan RJ. Action and valence modulate choice and choice-induced preference change. PLoS One. 2015; 10: e0119682. https://doi.org/10.1371/journal.pone.0119682 PMID: 25747703

54. Festinger L. A theory of cognitive dissonance. Stanford University Press; 1957.

55. Brehm JW. Postdecision changes in the desirability of alternatives. J Abnorm Psychol. 1956; 52: 384–389. https://doi.org/10.1037/h0041006 PMID: 13318848

56. Gerard HB, White GL. Post-Decisional Reevaluation of Choice Alternatives. Pers Soc Psychol Bull. 1983; 9: 365–369.

57. Steele CM, Spencer SJ, Lynch M. Self-image resilience and dissonance: the role of affirmational resources. J Pers Soc Psychol. 1993; 64: 885–896. https://doi.org/10.1037//0022-3514.64.6.885 PMID: 8326471

58. Heine SJ, Lehman DR. Culture, Dissonance, and Self-Affirmation. Pers Soc Psychol Bull. 1997; 23: 389–400.

59. Koriat A. The self-consistency model of subjective confidence. Psychol Rev. 2012; 119: 80–113. https://doi.org/10.1037/a0025648 PMID: 22022833

60. Zawadzka K, Higham PA, Hanczakowski M. Confidence in Forced-Choice Recognition: What Underlies the Ratings? J Exp Psychol Learn Mem Cogn. 2016. https://doi.org/10.1037/xlm0000321 PMID: 27685019

61. Körding KP, Tenenbaum JB. Causal inference in sensorimotor integration. NIPS. 2007.

62. Wei K, Körding KP. Causal Inference in Sensorimotor Learning. Sensory Cue Integration. 2011; 30–30.

63. Lau H. Consciousness, Metacognition, & Perceptual Reality Monitoring. PsyArxiv. 2019. https://doi.org/10.31234/osf.io/ckbyf

64. Merten K, Nieder A. Active encoding of decisions about stimulus absence in primate prefrontal cortex neurons. Proc Natl Acad Sci U S A. 2012; 109: 6289–6294. https://doi.org/10.1073/pnas.1121084109 PMID: 22460793

65. Fleming SM, Lau H. How to measure metacognition. Front Hum Neurosci. 2014; 8: 1–9. https://doi.org/10.3389/fnhum.2014.00001 PMID: 24474914

66. Fleming SM, Huijgen J, Dolan RJ. Prefrontal contributions to metacognition in perceptual decision making. J Neurosci. 2012; 32: 6117–6125. https://doi.org/10.1523/JNEUROSCI.6489-11.2012 PMID: 22553018

67. McCurdy LY, Maniscalco B, Metcalfe J, Liu KY, de Lange FP, Lau H. Anatomical coupling between distinct metacognitive systems for memory and visual perception. J Neurosci. 2013; 33: 1897–1906. https://doi.org/10.1523/JNEUROSCI.1890-12.2013 PMID: 23365229

68. Fleming SM, Weil RS, Nagy Z, Dolan R, Rees G. Relating Introspective Accuracy to Individual Differences in Brain Structure. Science. 2010;329. https://doi.org/10.1126/science.1197747 PMID: 20947752

69. Kao Y-C, Davis ES, Gabrieli JDE. Neural correlates of actual and predicted memory formation. Nat Neurosci. 2005; 8: 1776–1783. https://doi.org/10.1038/nn1595 PMID: 16286927

70. Janowsky JS, Shimamura AP, Kritchevsky M, Squire LR. Cognitive impairment following frontal lobe damage and its relevance to human amnesia. Behav Neurosci. 1989; 103: 548–560. https://doi.org/10.1037//0735-7044.103.3.548 PMID: 2736069

71. Pannu JK, Kaszniak AW. Metamemory experiments in neurological populations: a review. Neuropsychol Rev. 2005; 15: 105–130. https://doi.org/10.1007/s11065-005-7091-6 PMID: 16328731

72. Schnyer DM, Verfaellie M, Alexander MP, LaFleche G, Nicholls L, Kaszniak AW. A role for right medial prefontal cortex in accurate feeling-of-knowing judgements: evidence from patients with lesions to frontal cortex. Neuropsychologia. 2004; 42: 957–966. https://doi.org/10.1016/j.neuropsychologia.2003.11.020 PMID: 14998710

73. Lau H, Passingham RE. Relative blindsight in normal observers and the neural correlate of visual consciousness. Proceedings of the National Academy of Sciences. 2006; 103: 18763–18768. https://doi.org/10.1073/pnas.0607716103 PMID: 17124173

74. Middlebrooks PG, Sommer MA. Neuronal Correlates of Metacognition in Primate Frontal Cortex. Neuron. 2012; 75: 517–530. https://doi.org/10.1016/j.neuron.2012.05.028 PMID: 22884334

75. Lau H, Rosenthal D. Empirical support for higher-order theories of conscious awareness. Trends Cogn Sci. 2011; 15: 365–373. https://doi.org/10.1016/j.tics.2011.05.009 PMID: 21737339

76. Kim B, Basso M. Saccade target selection in the superior colliculus: a signal detection theory approach. J Neurosci. 2008; 28: 2991–3007. https://doi.org/10.1523/JNEUROSCI.5424-07.2008 PMID: 18354003

77. Smith PL, Ratcliff R. Psychology and neurobiology of simple decisions. Trends Neurosci. 2004; 27: 161–168. https://doi.org/10.1016/j.tins.2004.01.006 PMID: 15036882

78. Gold JI, Shadlen MN. Representation of a perceptual decision in developing oculomotor commands. Nature. 2000; 404: 390–394. https://doi.org/10.1038/35006062 PMID: 10746726

79. Horwitz GD, Batista AP, Newsome WT. Representation of an abstract perceptual decision in macaque superior colliculus. J Neurophysiol. 2004; 91: 2281–2296. https://doi.org/10.1152/jn.00872.2003 PMID: 14711971

80. Yeung N, Summerfield C. The Cognitive Neuroscience of Metacognition. 2014; 147–167.