



**HAL**  
open science

## Modeling perceptual confidence and the confidence forced-choice paradigm.

Pascal Mamassian, Vincent de Gardelle

► **To cite this version:**

Pascal Mamassian, Vincent de Gardelle. Modeling perceptual confidence and the confidence forced-choice paradigm.. *Psychological Review*, 2021, 129 (5), pp.976-998. 10.1037/rev0000312. hal-03329211

**HAL Id: hal-03329211**

**<https://hal.science/hal-03329211>**

Submitted on 6 Sep 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1                                   Modelling Perceptual Confidence and  
2                                   the Confidence Forced-Choice Paradigm

3                                   Pascal Mamassian<sup>1</sup> and Vincent de Gardelle<sup>2</sup>

4       (1) Laboratoire des systèmes perceptifs, Département d'études cognitives,  
5           École normale supérieure, PSL University, CNRS, Paris, France

6       (2) CNRS and Paris School of Economics, Paris, France

7    *Correspondence should be addressed to:*

8    Pascal Mamassian, Laboratoire des Systèmes Perceptifs (CNRS UMR 8248)

9    Ecole Normale Supérieure, 29 rue d'Ulm, 75005 Paris, France

10   Email: <pascal.mamassian@ens.fr>

11 **Abstract**

12 Perceptual confidence is an evaluation of the validity of our perceptual decisions. We present here  
13 a complete generative model that describes how confidence judgments result from some  
14 confidence evidence. The model that generates confidence evidence has two main parameters,  
15 confidence noise and confidence boost. Confidence noise reduces the sensitivity to the confidence  
16 evidence, and confidence boost accounts for information used for confidence judgment which was  
17 not used for the perceptual decision. The opposite effect of these two parameters creates  
18 confidence metamers, where the confidence in a perceptual decision is the same in spite of  
19 differences in confidence noise and confidence boost. When the data set is rich enough, both of  
20 these parameters can be recovered, thus allowing us to estimate the extent to which confidence is  
21 generated in parallel or serially to the perceptual decision. We also describe a novel measure of  
22 confidence efficiency relative to the ideal confidence observer, as well as the estimate of one type  
23 of confidence bias. Finally, we apply the model to the confidence forced-choice paradigm, a  
24 paradigm that provides objective estimates of confidence, and we discuss how each parameter of  
25 the model can be recovered using this paradigm.

26 *Keywords:* meta-perception, visual confidence, modelling, efficiency, confidence forced-choice

## 27 1. Introduction

28 Metacognition is the ability of individuals to monitor and regulate their own cognitive processes  
29 (Nelson & Narens, 1990). Therefore, in the case of perception, metaperception is the ability of  
30 individuals to monitor and control their perceptual decisions (Mamassian, 2020). When making a  
31 choice, a key expression of metacognition is the confidence associated with the decision. Correctly  
32 inferring our own level of performance is clearly important for an individual, as confidence might be  
33 used to regulate learning (e.g. Hainguerlot et al., 2018), allocate resources to a particular task (e.g.  
34 van den Berg et al., 2016), compare different tasks (de Gardelle & Mamassian, 2014) and prioritize  
35 them (Aguilar-Lleyda et al., 2020). Perceptual confidence, and more broadly metacognition, has  
36 been extensively reviewed elsewhere (e.g. Fleming et al., 2012; Yeung & Summerfield, 2012;  
37 Meyniel et al., 2015; Mamassian, 2016; Pouget et al., 2016).

38 One issue of primary importance in metaperception is whether confidence judgments are based on  
39 the same information as that used for the perceptual decisions. Even though confidence is an  
40 evaluation of the validity of our perceptual decisions, it is plausible that the computation of  
41 confidence involves some information that is processed in parallel to (e.g. Fleming & Daw, 2017)  
42 or after (e.g. Pleskac & Busemeyer, 2010) the perceptual decision. The difficulty in establishing the  
43 extent to which confidence is processed along a parallel stream of information is that there are  
44 other factors that affect the quality of confidence judgments. In particular, the computation of  
45 confidence might rest on degraded perceptual information (e.g. Bang et al., 2019). Therefore, it is  
46 important to have a good theoretical framework within which the different factors that contribute to  
47 confidence are clearly defined.

48 There are currently two main frameworks used for the study of confidence, one based on Signal  
49 Detection Theory (SDT), and the other based on evidence accumulation (for a review, see  
50 Mamassian, 2016). The SDT framework (Green & Swets, 1966) has been exceedingly successful  
51 for modelling choice tasks, also referred to as Type 1 tasks, and it also formed the basis for  
52 discussing confidence judgments, also known as Type 2 judgments (Clarke et al., 1959; Galvin et  
53 al., 2003). However, this framework is silent about *how* Type 2 judgments are actually made. The  
54 primary aim of the present manuscript is to provide a complete generative model for perceptual  
55 confidence judgments that is grounded in SDT. With this generative model, we have three main  
56 objectives that we briefly introduce next. These objectives focus on the separation of serial and  
57 parallel processing of confidence, a measure of confidence efficiency that is defined at the  
58 metacognitive level, and an estimate of one critical form of confidence bias.

59 Our model of confidence is based on the idea that confidence judgments are based on the current  
60 perceptual decision and some decision variable that we call *confidence evidence*. Confidence  
61 evidence is obtained from two possible streams of information processing. Through the serial

62 stream, confidence evidence is just a duplicate of the sensory evidence that is used for the  
63 perceptual decision, albeit, with additional sources of inefficiencies to duplicate this sensory  
64 information. This stream of processing is present in all models of confidence. In contrast, through  
65 the parallel stream, confidence evidence has novel access to the physical stimulus, independently  
66 from the processing that led to the perceptual decision. Note that a similar distinction between  
67 hierarchical and dual-channel models can be found in other theoretical frameworks (e.g.  
68 Maniscalco & Lau, 2016). The first objective of our modelling effort is thus to clarify the respective  
69 contributions of the serial and parallel streams to confidence judgments, both theoretically and  
70 empirically.

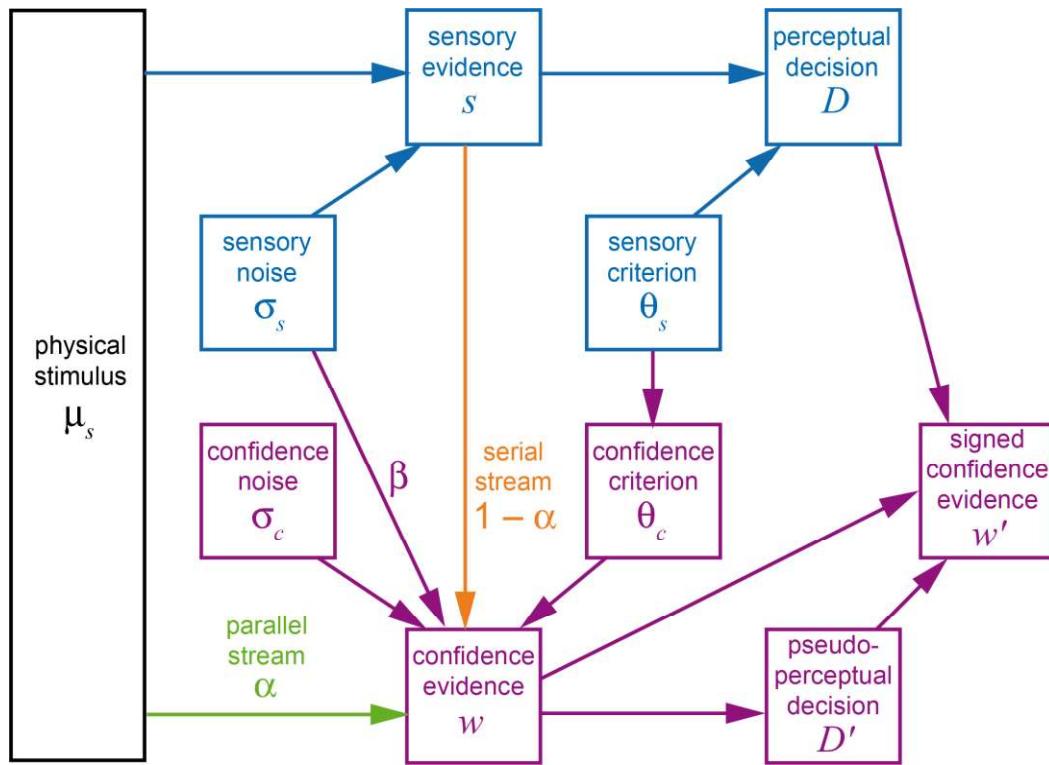
71 Figure 1 illustrates our modelling approach and provides the links between the different variables  
72 of the model. All the notations of the model are summarized in Table 1. We highlight in particular  
73 two components in relation to the serial and parallel streams of processing. The first component is  
74 the *confidence noise* which characterizes the inefficiency of the confidence evidence computation  
75 relative to the *ideal confidence observer*. The second component is the *confidence boost* which  
76 characterises the relative contribution of the parallel stream to confidence evidence. The reason  
77 why this latter component is called confidence boost is because new evidence from the stimulus  
78 will augment the information present at the Type 2 level and boost metacognitive efficiency  
79 towards a *super-ideal* level.

80 Confidence boost and confidence noise have opposite effects on Type 2 performance, and it is  
81 difficult to properly estimate both of them in practice. Yet, it is important to have at our disposal an  
82 overall measure of Type 2 efficiency. Defining such a measure has been challenging in the past  
83 (Fleming & Lau, 2014), but a significant step forward was obtained thanks to the meta-d'  
84 computation recently (Maniscalco & Lau, 2012). This methodological tool allows experimenters to  
85 measure metacognitive abilities without confounds from Type 1 performance. However, one key  
86 characteristic of this measure is that it uses the metric of the Type 1 task, rather than of the Type 2  
87 task. The second objective of our modelling effort is thus to offer a measure of Type 2 efficiency  
88 that is really anchored to the Type 2 level of processing.

89 The third objective of our modelling effort is to be able to detect some confidence biases. In our  
90 model, we focus on one particular type of confidence biases, where an over-confidence represents  
91 an over-estimation of one's perceptual sensitivity, or equivalently an under-estimate of the sensory  
92 noise. This type of confidence biases is difficult to detect because all the confidence judgments for  
93 a particular task are affected. When confidence is compared across two distinct tasks, we can  
94 obtain an estimate of the over-confidence for one task relative to the other. This kind of confidence  
95 comparison forms the basis of the confidence forced-choice paradigm. In this procedure,  
96 participants complete two Type 1 decisions on distinct stimuli, and then indicate which decision  
97 was associated with the greater confidence (Barthelmé & Mamassian, 2009; de Gardelle &

98 Mamassian, 2015). We apply our generative model to the confidence forced-choice paradigm and  
 99 discuss how reliably each parameter of the model can be estimated in this paradigm.

100



101

102 Figure 1. Overall framework for perceptual and confidence decision making. For Type  
 103 1 processing (in blue), the perceptual decision is based on sensory evidence that is an  
 104 estimate of the physical stimulus. Sensory evidence is corrupted by sensory noise. For  
 105 Type 2 processing (in purple), the confidence judgment is based on confidence  
 106 evidence that is a combination of information processed in serial (orange) and parallel  
 107 (green) streams. The serial stream duplicates the sensory evidence whereas the  
 108 parallel stream allows for another look at the physical stimulus. Confidence evidence  
 109 is corrupted by confidence noise. It is also normalized by an estimate of sensory noise  
 110 that is possibly corrupted by a confidence bias, and it is compared to a confidence  
 111 criterion that possibly differs from the sensory criterion. Finally, the signed confidence  
 112 evidence is the magnitude of the confidence evidence that acquires a negative sign if  
 113 the perceptual decision is incompatible with confidence evidence. See text for details.

114

115 As we compute confidence efficiency, we will see that the same confidence efficiency level can be  
 116 achieved as a trade-off between confidence noise and confidence boost. The values of confidence

117 noise and boost which give rise to the same confidence efficiency form a family that we call  
 118 confidence metamers.

119

Notation	Meaning	Domain
$\mu_s$	Stimulus strength	$(-\infty, +\infty)$
$s$	Sensory evidence	$(-\infty, +\infty)$
$w$	Confidence evidence	$(-\infty, +\infty)$
$w'$	Signed confidence evidence	$(-\infty, +\infty)$
$\sigma_s$	Sensory noise (standard deviation of normal distribution) that drives perceptual sensitivity	$[0, +\infty)$
$\theta_s$	Sensory criterion that drives bias in the perceptual decision	$(-\infty, +\infty)$
$D$	Perceptual decision based on sensory evidence	
$D'$	Pseudo perceptual decisions based on confidence evidence	
$C$	Confidence choice, i.e. interval chosen as more confident with respect to the self-consistency of the perceptual decision	$\{1, 2\}$
$\sigma_c$	Confidence noise (standard deviation)	$[0, +\infty)$
$\theta_c$	Confidence criterion against which confidence evidence is evaluated	$(-\infty, +\infty)$
$\alpha$	Confidence boost, i.e. the fraction of super-ideal confidence performance	$[0, 1]$
$\beta$	Confidence bias in over-estimating one's sensory sensitivity	$(0, +\infty)$
$\gamma$	Interval bias in favour of interval 1 in a confidence pair	$(-\infty, +\infty)$
$F(s_1, s_2)$	Joint distribution of sensory evidence in confidence pair	
$G(w_1, w_2   s_1, s_2)$	Joint distribution of confidence evidence $w$ conditional on sensory evidence $s$ in confidence pair	
$H(s, w)$	Joint distribution of sensory and confidence evidence (its covariance matrix is $K$ )	
$Q(s; \mu_s, \sigma_s)$	Mean of the distribution of confidence evidence conditional on a particular value of sensory evidence $s$	$(-\infty, +\infty)$
$\tau$	Equivalent confidence noise (standard deviation)	$[0, +\infty)$
$\eta$	Confidence efficiency	$[0, +\infty)$

120

121 Table 1. Notations used in this manuscript.

122

123 Our manuscript is organized as follows. In the next two sections, we define what we mean by  
124 confidence in this manuscript, and then review briefly the confidence forced-choice paradigm. In  
125 section 4, we define the confidence ideal and super-ideal observers, which will help us determining  
126 the different ways confidence computation can be inefficient. We then detail our generative model  
127 in sections 5 and 6, describing how confidence evidence is linked to sensory evidence, and in  
128 sections 7 and 8, we apply this model to the confidence forced-choice paradigm. Section 9  
129 introduces the notion of confidence metamers and explains how confidence efficiency is computed.  
130 We finish by showing the robustness of the parameter estimation (section 10), including the  
131 confidence bias (section 11), and illustrate in section 12 how the model can be fitted to real data by  
132 re-analysing one of our previous studies. Finally, section 13 presents a discussion of our approach.

## 133 **2. Defining Confidence as Subjective Self-Consistency**

134 We start by formally defining confidence in a perceptual decision as the subjective estimation  
135 made by an observer that her decision is self-consistent. Here, self-consistency refers to an  
136 agreement between the current perceptual decision and the most frequent decision made by the  
137 observer for a given stimulus and experimental conditions. Perceptual confidence is thus an  
138 estimation of the probability that the same decision would be made again, given the same physical  
139 stimulus and experimental conditions. In terms of Signal Detection Theory (SDT), self-consistency  
140 relates to perceptual sensitivity, disregarding perceptual bias.

141 Note that our definition slightly departs from the classic definition of confidence as an estimate of  
142 perceptual accuracy (i.e. probability of being correct). The difference between the two definitions is  
143 best illustrated by considering cases of perceptual illusions due to a sensory bias. In such cases,  
144 observers can be consistently incorrect in their decisions but still relatively confident in their  
145 perception. By focusing on self-consistency, rather than accuracy, our definition does not force us  
146 to call all observers overconfident in this case, which may be desirable given that the bias arises  
147 here at the perceptual level and not at the metacognitive level *per se*. If we follow the classic  
148 definition of confidence, however, we would have to conclude that the observer is overconfident  
149 because she is both incorrect and very confident.

150 Our definition of confidence as an estimate of one's own self-consistency aligns with other works.  
151 In meta-memory, Koriat (2012) has highlighted that confidence may reflect the consensuality of  
152 one's own answer with respect to answers chosen by other individuals, rather than just whether  
153 one's answer is correct or not. Our discussion of overconfidence is also reminiscent of one  
154 particular type of overconfidence discussed in the literature. Three types of overconfidence are  
155 sometimes distinguished, namely the *overestimation* of one's accuracy, the *overplacement* relative  
156 to others, and the *overprecision* of one's beliefs (Moore & Healy, 2008). Our definition of



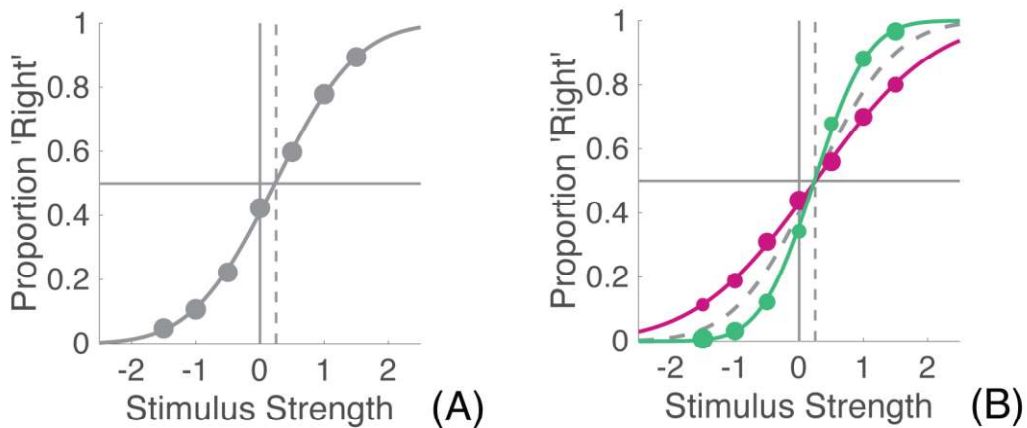
157 confidence as subjective self-consistency naturally fits with overprecision. In other words, with our  
158 definition, an individual would be overconfident in a perceptual task if she overestimates her own  
159 sensitivity in this task. By contrast, the traditional definition of confidence as the subjective  
160 probability of being correct corresponds to overconfidence being an overestimation of one's  
161 accuracy. Note that in the SDT framework, these two definitions would be equivalent if all decision  
162 criteria are neutral. However, as detailed below, our modelling approach will allow for any criteria,  
163 including criteria that differ between Type 1 and Type 2 evaluation of the evidence.

### 164 **3. Confidence Forced-Choice**

165 In this manuscript, we focus on the confidence forced-choice paradigm. One key advantage of this  
166 procedure is to bypass the rating scale typically used to measure confidence, and to focus directly  
167 on the internal confidence, eliminating the need for participants to maintain a constant mapping  
168 between internal confidence and ratings. In this paradigm, participants indicate which of two  
169 intervals produces the highest feeling of confidence, where each interval consists of a stimulus,  
170 and a decision made on that stimulus. A *confidence trial* is thus composed of two stimuli, two  
171 perceptual decisions, and the confidence comparison choice between these two decisions.

172 Let us consider a typical use of the confidence forced-choice paradigm around a psychophysical  
173 experiment. In this example, the perceptual task is to indicate whether the dots of a random-dot  
174 kinematogram stimulus are moving to the right or to the left relative to a reference direction. Stimuli  
175 differ in strength, manipulated from trial to trial in how much the motion direction deviates from the  
176 reference. Stimulus strength affects how well observers can discriminate the direction of motion, as  
177 represented by the psychometric function (Figure 2A). The slope of the psychometric function  
178 reflects the sensitivity of the observer in the perceptual task.

179 To examine how confidence relates to perceptual sensitivity, we can analyse separately the  
180 perceptual decisions associated with higher and lower confidence in each confidence trial. We can  
181 then replot the psychometric function separately for these *confidence-chosen* and for *confidence-*  
182 *declined* decisions (Figure 2B). In the example of the figure, these two new psychometric functions  
183 are distinct, the one for the confidence-chosen decisions presents a steeper slope than the one for  
184 the confidence-declined decisions, or than the original one estimated over all trials (Figure 2A).  
185 This property is a signature of meta-perception, as it indicates that participants were able to pick  
186 the interval that led to a better performance, at least for some trials. If the participants gave their  
187 metacognitive judgments at random, as if they were not able to judge the quality of their perceptual  
188 decisions, the psychometric functions for chosen and declined decisions would overlap completely.  
189 In contrast, when the observer is using all the information she can use for her confidence  
190 judgment, the gain in the slope of the psychometric functions is strictly larger than zero.



192

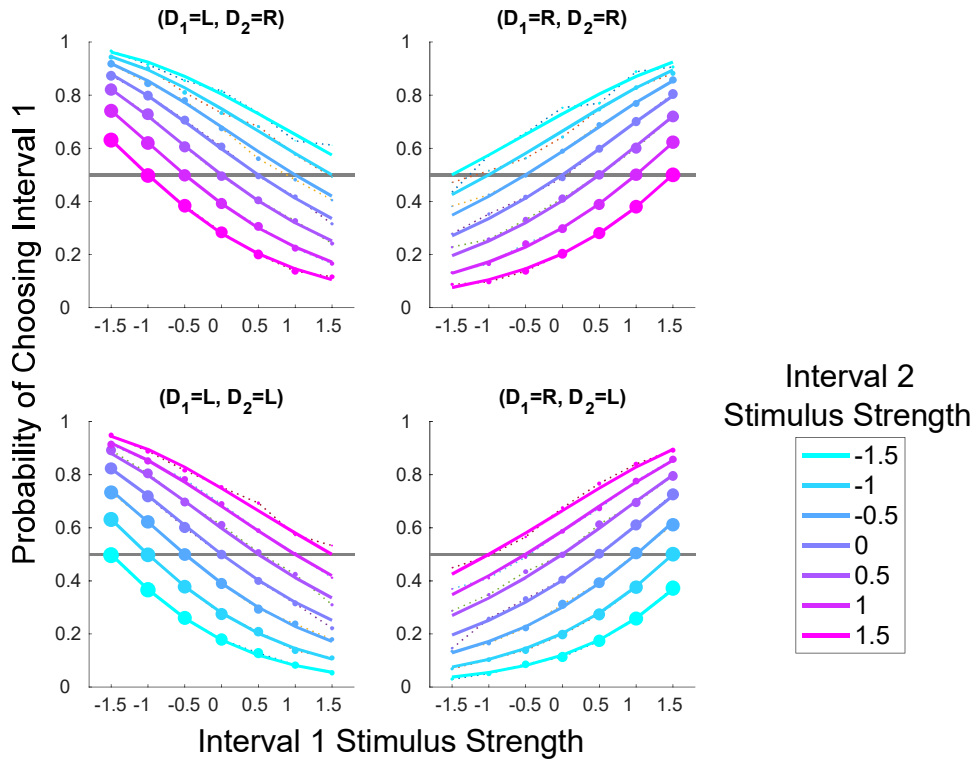
193 Figure 2. Psychometric functions. (A) Original psychometric function. The  
 194 psychometric function links stimulus strength to perceptual decision, here the  
 195 proportion of dots moving rightward. The solid line is a cumulative Gaussian fit to the  
 196 psychometric functions. The standard deviation of the best fit determines its slope  
 197 (here 1.01, a good approximation of the parameter  $1/\sigma_s$  used in the simulation).  
 198 (B) Psychometric function split by confidence. Trials judged to have higher confidence  
 199 are sorted out and a new psychometric function is plotted for these trials only (green  
 200 points). The remaining trials have been declined for confidence (red points). Dots size  
 201 is proportional to the number of trials in this condition. For the psychometric function  
 202 based on the chosen trials for confidence, the best fit gives a slope of 1.47. The gain in  
 203 the slope of the psychometric functions from the unsorted (grey dashed curve) to the  
 204 chosen (green curve) trials is therefore  $1.47/1.01 = 1.45$ . The parameters used to  
 205 generate this and the following figures are provided in Table 2.

206

207 Even though it is simple and natural to use the gain in the slope of psychometric functions as an  
 208 index of metacognitive ability (see, e.g. Barthelmé & Mamassian, 2009; De Martino et al., 2013; de  
 209 Gardelle & Mamassian, 2014, 2015), we introduce later the confidence efficiency as an alternative  
 210 descriptor of confidence sensitivity. Indeed, the comparison of psychometric functions actually  
 211 discards important information about which confidence pairs were presented to participants. The  
 212 full data set includes not only how a given perceptual trial falls into the confidence-chosen or  
 213 confidence-declined set, but also how the confidence comparison choice depends on the two trials  
 214 within a pair, which may have different stimulus strengths and different decisions. In the example  
 215 of the simulated experiment shown in Figure 2, there were 7 possible stimulus strengths and two  
 216 possible perceptual decisions ('R' or 'L') for each interval in a confidence pair, leading to 196 ( $7 \times 7$   
 217  $\times 2 \times 2$ ) possible combinations. In each of these combinations, we can measure the probability

218 that interval 1 is associated with a greater confidence than interval 2. These confidence choice  
219 probabilities are illustrated in Figure 3.

220



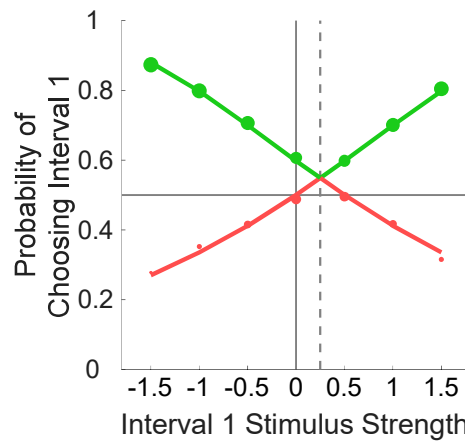
221

222 Figure 3. Confidence choice probabilities for each combination of stimulus strengths.  
223 Each panel shows the probability of choosing the first interval as the more confident  
224 one given the stimulus strength presented in the first interval (x-axis) and in the  
225 second interval (coloured lines). The different panels correspond to the four different  
226 pairs of perceptual decisions across the two intervals (e.g. responses  $D_1 = 'L'$  and  
227  $D_2 = 'R'$  in the top left panel). Dot size is proportional to the number of trials obtained in  
228 the simulation for this particular combination of stimulus strengths and Type 1  
229 responses. Dotted lines link points that have the same stimulus strength in the second  
230 interval. The solid curves show the best fitted model described later in the manuscript.  
231 In this plot, parameters are those listed in Table 2, except  $n = 100,000$ .

232

233 From the simulations shown in Figure 3, we see that confidence depends on the interaction  
234 between stimulus strength and perceptual decision, as typically found in empirical data. To better  
235 illustrate this pattern, let us focus on one subset where the stimulus strength in the second interval  
236 is 0 and the perceptual decision for this stimulus is 'R'. This subset corresponds to the middle blue  
237 line in the two top panels, which are replotted on Figure 4 but in different colours. Specifically, self-

238 consistent perceptual decisions are shown in green, and self-inconsistent decisions in red. Here,  
239 self-consistent decisions correspond to responding 'R' for stimulus strengths in the first interval that  
240 are above the sensory criterion (0.25), and responding 'L' below. As expected, the probability of  
241 choosing the first interval with greater confidence is always larger for self-consistent than for self-  
242 inconsistent decisions. In addition, as stimulus strength deviates more from the sensory criterion,  
243 confidence increases for self-consistent decisions, and decreases for self-inconsistent decisions.  
244 This is expected from a participant who displays meta-perception, although the exact form of this  
245 X-pattern varies across experimental conditions and models of confidence (Sanders et al., 2016;  
246 Adler & Ma, 2018; Rausch & Zehetleitner, 2019).



247

248 Figure 4. Choice probabilities for self-consistent and inconsistent decisions. The plot  
249 shows the same data as in Figure 3, for one particular sensory stimulus and  
250 perceptual decision in interval 2. Self-consistent perceptual decisions are shown in  
251 green, and self-inconsistent decisions in red. The probability of choosing the first  
252 interval with greater confidence is larger for self-consistent than for self-inconsistent  
253 decisions. In addition, as stimulus strength deviates more from the sensory criterion,  
254 confidence increases for self-consistent decisions, and decreases for self-inconsistent  
255 decisions. The resulting X-pattern is symmetric about the vertical axis passing through  
256 the sensory criterion.

257

258 We are now interested in modelling the sensitivity with which participants can estimate their  
259 confidence in their perceptual decisions. The model will attempt to replicate all 196 different  
260 probabilities that interval 1 is the winner of the confidence decision in Figure 3. In particular, we are  
261 interested in describing the ideal confidence observer that is using the exact same information for  
262 confidence judgments as the perceptual decisions, so that we can compare human meta-  
263 perceptual sensitivity to this ideal confidence observer. Along the way, we will also define a super-

264 ideal confidence observer that maximizes confidence performance. Unless otherwise noted, the  
 265 parameters in the figures take the default values shown in Table 2.

266

Parameter	Meaning	Figure Value	Ideal Value
$\{\mu_A, \mu_B\}$	Examples of stimulus strengths	$\{1.5, -0.5\}$	
$\mu_s$	Stimulus strengths for a complete simulated experiment	$-1.5:0.5:1.5$	
$(s_1, s_2)$	Sensory evidence in intervals 1 and 2 of a confidence pair where stimulus strengths are $(\mu_A, \mu_B)$	$(0.9, 0.7)$	
$(D_1, D_2)$	Perceptual decisions in intervals 1 and 2 of a confidence pair where stimulus strengths are $(\mu_A, \mu_B)$	$(R, R)$	$(R, L)$
$\sigma_s$	Sensory noise (standard deviation)	1.0	0.0
$\theta_s$	Sensory criterion that drives bias in the perceptual decision	0.25	0.0
$\sigma_c$	Confidence noise (standard deviation)	0.5	0.0
$\theta_c$	Confidence criterion	0.0	0.0
$\alpha$	Confidence boost	0.2	0.0
$\beta$	Confidence bias in over-estimating one's sensory sensitivity	1.0	1.0
$\gamma$	Interval bias in favour of interval 1 in a confidence pair	0.0	0.0
$n$	Number of confidence pairs in a simulation	10,000	

267

268 Table 2. Unless explicitly stated in the figure caption, the parameter values used in the  
 269 figures are the ones in this table. In the last column are shown the values  
 270 corresponding to the ideal observer and ideal confidence observer.

271

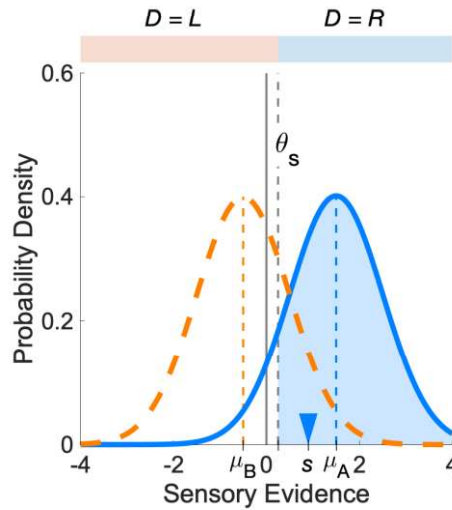
#### 272 4. Ideal Confidence Observer

273 In this section, we present how the perceptual decision is derived from sensory evidence. By  
 274 analogy, we introduce the confidence evidence that will be the basis for the confidence judgment.  
 275 The approach is based on Signal Detection Theory (Green and Swets, 1966) and ideal observer  
 276 principles (Barlow, 1962; Geisler, 1989). In the next section, we will generalize this model of  
 277 confidence by considering several ways in which actual confidence judgments can deviate from  
 278 optimal ones.

279 4.a. Perceptual Decisions

280 We consider here a perceptual task in which a stimulus has to be categorized as ‘Right’ (‘R’) or  
 281 ‘Left’ (‘L’). In a typical psychophysical experiment, there will be a range of stimuli with different  
 282 levels of difficulty that we represent by the stimulus strength  $\mu_s$ . For illustrative purposes, we first  
 283 consider two such stimuli, A and B, that belong to categories ‘R’ and ‘L’ respectively (Figure 5).

284



285

286 Figure 5. Sensory evidence in a perceptual discrimination task. Stimuli to be  
 287 discriminated belong to two categories ‘Right’ (‘R’) and ‘Left’ (‘L’). The distribution of  
 288 sensory evidence for two stimuli A and B is in blue and orange, respectively. On each  
 289 trial, the participant has access to one sample of the stimulus category presented on  
 290 that trial (a sample  $s$  from stimulus A is shown by the blue triangle). All sensory  
 291 evidence to the right of sensory criterion  $\theta_s$ , represented by the blue shaded area, are  
 292 assigned to the ‘R’ category.

293

294 Because of sensory noise, the observer only has access to some noisy sensory evidence  $s$ . We  
 295 assume that on average the observer has an unbiased estimate of the sensory strength, so the  
 296 mean of  $s$  is  $\mu_s$ . For simplicity, we further assume that the sensory noise is normally distributed,  
 297 with common variance  $\sigma_s^2$  for all stimuli, such that a sensory noise sample  $\epsilon_s$  for one particular trial  
 298 follows the distribution  $\epsilon_s \sim N(0, \sigma_s^2)$ .

299 The sensory evidence on one trial is then

300 
$$s = \mu_s + \epsilon_s , \tag{1}$$

301 where  $\mu_s = \mu_A$  if stimulus A was presented, and  $\mu_s = \mu_B$  if stimulus B was presented instead. A  
302 perceptual decision (Type 1 decision  $D$ ) consists in comparing the sensory evidence against a  
303 sensory criterion  $\theta_s$ , namely

$$304 \quad \begin{cases} D = \text{'R'} & \text{if } s > \theta_s, \\ D = \text{'L'} & \text{otherwise} \end{cases} . \quad (2)$$

305 The most frequent percept for stimulus A is 'R' (the blue shaded area in Figure 5 to the right of the  
306 sensory criterion is larger than 0.5 because  $\mu_A > \theta_s$ ). Therefore, when stimulus A is presented, the  
307 perceptual decision will be self-consistent if it is 'R'. We present other properties of self-  
308 consistency in Appendix A.

#### 309 4.b. Ideal Confidence Observer

310 Now that we have modelled perceptual decisions, we can consider confidence (Type 2) judgments.  
311 We start with the important case of the *ideal confidence observer* that will be used as a reference  
312 to compare human confidence judgments. The ideal confidence observer is ideal for its confidence  
313 judgment but suboptimal for its perceptual decision. In other words, this particular observer has the  
314 same sensory sensitivity and biases as the human observer, and thus is similarly subject to  
315 sensory noise and sensory criterion shifts as the human observer. However, it is ideal in the sense  
316 that it is able to judge optimally which of two perceptual decisions is more likely to be self-  
317 consistent based on the same sensory information that has been used to reach the perceptual  
318 decisions. In other words, for the ideal confidence observer, the confidence evidence will be  
319 entirely determined by the sensory evidence.

320 From Figure 5, we see that the perceptual decision is more likely to be self-consistent when the  
321 sensory evidence  $s$  is further away from the sensory criterion  $\theta_s$  (for a formal description of the  
322 probability of being self-consistent, see Appendix A). Therefore, from the point of view of the ideal  
323 confidence observer, the distance of the sensory evidence to the perceptual decision boundary is a  
324 good decision variable to estimate confidence (Galvin et al., 2003). We follow this tradition with  
325 one particular twist. To be able to estimate confidence sensitivity irrespective of the sensory  
326 sensitivity of the observer for the current task, we normalize the distance to the decision boundary  
327 by the sensory noise. As can be seen in Appendix D, this step alleviates apparent contradictions  
328 such that sensory noise increases metacognitive efficiency (Bang et al., 2019). In summary, we  
329 define the ideal confidence evidence to be

$$330 \quad w_{\text{ideal}} = (s - \theta_s) / \sigma_s . \quad (3)$$

331 Because confidence evidence has been normalized by sensory noise, it is a unit-free measure of  
332 confidence. In other words, it is not bound to the stimulus dimension that is relevant for a task (e.g.  
333 the angle in degrees of motion direction if the task of the observer is to estimate motion direction).  
334 This property is useful when comparing confidence across tasks (de Gardelle & Mamassian,  
335 2014). Further motivation for this choice of ideal confidence evidence is presented in Appendix A.

#### 336 4.c. Super-Ideal Confidence Observer

337 In contrast to the ideal confidence observer, the *super-ideal confidence observer* has access to the  
338 original stimulus, and not just the noisy sensory evidence used to make the perceptual decision.  
339 This scenario can actually lead to better performance than the ideal confidence observer, thus the  
340 term “super-ideal” confidence observer. This extreme scenario is interesting to consider because  
341 confidence judgments are often performed after perceptual decisions, and thus can benefit from a  
342 more extensive analysis (e.g. Pleskac & Busemeyer, 2010) or second look at the stimulus.  
343 Confidence evidence for the super-ideal confidence observer is now

$$344 \quad w_{\text{super\_ideal}} = (\mu_s - \theta_s) / \sigma_s . \quad (4)$$

345 Note that we still normalize the stimulus strength  $\mu_s$  relative to the sensory noise  $\sigma_s$  and sensory  
346 criterion  $\theta_s$  so as to obtain a unit-free measure of confidence that still reflects the potential  
347 perceptual bias of the observer.

### 348 **5. Generative Model of Confidence Evidence**

349 In the previous section, we have described the ideal and super-ideal confidence observers. We  
350 now consider four ways in which human confidence judgments can deviate from the ideal  
351 confidence observer. First, human observers can behave partially as the super-ideal confidence  
352 observer, thereby boosting their confidence sensitivity. Second, they can display some confidence  
353 noise that is impairing their ability to use their confidence evidence. Third, human observers can be  
354 inaccurate in their estimate of the sensory sensitivity, thereby generating over- or under-  
355 confidence. Finally, human observers can be inaccurate in their estimate of the sensory bias,  
356 thereby creating potential conflicts between sensory and confidence decisions. We now examine  
357 these four cases in turn.



358 5.a. Confidence Boost

359 We define *confidence boost*, noted  $\alpha$ , the fraction of the super-ideal confidence observer that  
360 contributes to the human confidence evidence. If  $\alpha = 1$ , then the human observer is just like the  
361 super-ideal confidence observer, and if  $\alpha = 0$ , then the human observer behaves just like the ideal  
362 confidence observer. Confidence evidence now becomes a mixture of the evidence from the  
363 super-ideal and ideal confidence observers, namely

$$364 \quad w = \alpha \cdot w_{\text{super\_ideal}} + (1 - \alpha) \cdot w_{\text{ideal}} . \quad (5)$$

365 This expression can be rewritten as

$$366 \quad w = (\alpha \cdot \mu_s + (1 - \alpha) \cdot s - \theta_s) / \sigma_s$$
$$367 \quad w = (\mu_s + (1 - \alpha) \cdot \epsilon_s - \theta_s) / \sigma_s . \quad (6)$$

368 The effect of confidence boost on the psychometric function is shown in Figure 6A. This  
369 psychometric function should be compared to the one with the default parameters in Figure 2B.  
370 When confidence boost increases, we observe a steeper psychometric function for the confidence-  
371 chosen trials. In other words, the observer is better able to discriminate correct from incorrect  
372 perceptual decisions. This is not surprising as the confidence boost reflects the ability of the  
373 observer to use more information at the metacognitive level.

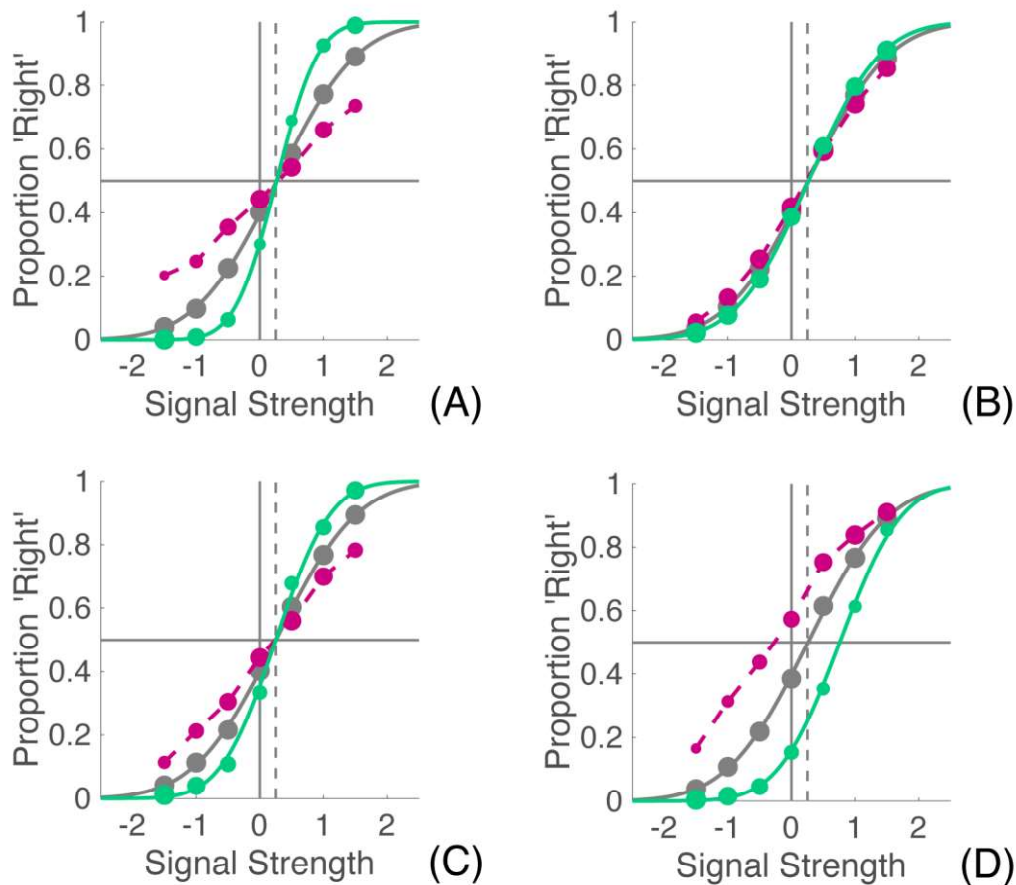
374 5.b. Confidence Noise

375 Just like sensory noise corrupts the sensory evidence, we introduce *confidence noise* that corrupts  
376 the confidence evidence. We model confidence noise as a zero-mean normal distribution with  
377 variance  $\sigma_c^2$ , such that a confidence noise sample  $\epsilon_c$  follows the distribution  $\epsilon_c \sim N(0, \sigma_c^2)$ . We  
378 assume that confidence noise is additive and independent of sensory evidence, so the new  
379 confidence evidence becomes

$$380 \quad w = (\mu_s + (1 - \alpha) \cdot \epsilon_s - \theta_s) / \sigma_s + \epsilon_c . \quad (7)$$

381 Because confidence noise is unrelated to the sensory evidence, it is unit-less, and comparable  
382 across different tasks (see e.g. de Gardelle & Mamassian, 2014). The effect of confidence noise  
383 on the psychometric function is shown in Figure 6B. When the confidence noise increases, we  
384 obtain a shallower psychometric function for the confidence chosen trials. In other words, the  
385 observer is less able to discriminate correct from incorrect perceptual decisions.

386



387

388 Figure 6. Influence of different model parameters on the psychometric functions. In  
 389 these plots, the parameters are those listed in Table 2, except for one parameter. (A)  
 390 The confidence boost is increased to  $\alpha = 0.8$ . (B) The confidence noise is increased  
 391 to  $\sigma_c = 2.0$ . (C). The confidence bias is increased to  $\beta = 2.0$ . (D). The confidence  
 392 criterion is increased to  $\theta_c = 1.0$ .

393

394 5.c. Confidence Bias

395 Sensory evidence needs to be scaled to generate the confidence evidence such that the latter is  
 396 task-independent and unit-free. This is achieved by normalizing confidence evidence relative to the  
 397 sensory sensitivity, and consequently, confidence evidence is a good proxy for the probability of  
 398 being self-consistent in the perceptual decision (see again Appendix A). From the ideal confidence  
 399 observer perspective, this scaling factor should be the inverse of the sensory noise ( $1/\sigma_s$ ). We  
 400 represent by  $\beta$  the *confidence bias* which stands as a deviation away from this ideal scaling (this  
 401 corresponds to replacing  $1/\sigma_s$  with  $\beta/\sigma_s$ ). Values of  $\beta$  larger than 1.0 indicate over-confidence, and

402 values smaller than 1.0 under-confidence. Taking into account this misestimate of the sensory  
403 sensitivity leads to a new confidence evidence

$$404 \quad w = (\mu_s + (1 - \alpha) \cdot \epsilon_s - \theta_s) \cdot \beta / \sigma_s + \epsilon_c \quad (8)$$

405 The effect of confidence bias on the psychometric function is shown in Figure 6C. We observe that  
406 the psychometric function for the confidence chosen trials is not affected by the confidence bias  
407 (Figure 6C is identical to Figure 2B). This is not surprising since this parameter scales the  
408 confidence evidence in both intervals in the same way. Even though the effects of confidence bias  
409 are invisible here, we present below a condition where this confidence bias can be partially  
410 estimated (see section 11).

#### 411 5.d. Confidence Criterion

412 Finally, human observers can use a criterion against which they measure their confidence that is  
413 distinct from the sensory criterion. We represent by  $\theta_c$  the deviation of the *confidence criterion*  
414 away from the sensory criterion. Ideally this parameter is zero ( $\theta_c = 0$ ), but when it is not, the  
415 confidence evidence becomes

$$416 \quad w = (\mu_s + (1 - \alpha) \cdot \epsilon_s - \theta_s - \theta_c) \cdot \beta / \sigma_s + \epsilon_c \quad (9)$$

417 The effect of confidence criterion on the psychometric function is shown in Figure 6D. When the  
418 confidence criterion deviates from the sensory criterion, the point of subjective equality (PSE) for  
419 the confidence-chosen decisions (green curve) becomes different from the PSE for the original  
420 psychometric function (grey curve). The shift in PSE is coming from the inconsistency between the  
421 perceptual decision and what we will call the “pseudo perceptual decision” (see section 6.c), for a  
422 range of sensory values near the sensory criterion.

## 423 **6. Covariation of Sensory and Confidence Evidence**

424 Because of noise at the perceptual level or at the confidence level, sensory evidence and  
425 confidence evidence will vary across trials, even when the stimuli and the responses are the same.  
426 We will now characterise this variation, by defining the joint distribution of sensory and confidence  
427 evidence. This will allow us to produce summary statistics that will be useful for presenting the full  
428 model of the confidence comparison task. We note that previous models of confidence have  
429 discussed the joint distribution between sensory and confidence evidence (Fleming & Daw, 2017).  
430 However, it is important to appreciate that our definition is different from these previous studies

431 because our joint distribution is derived from a generative model based on the introduction of  
 432 confidence noise and confidence boost instead of being an arbitrary bivariate distribution function.

433 6.a. Joint distribution for sensory and confidence evidence

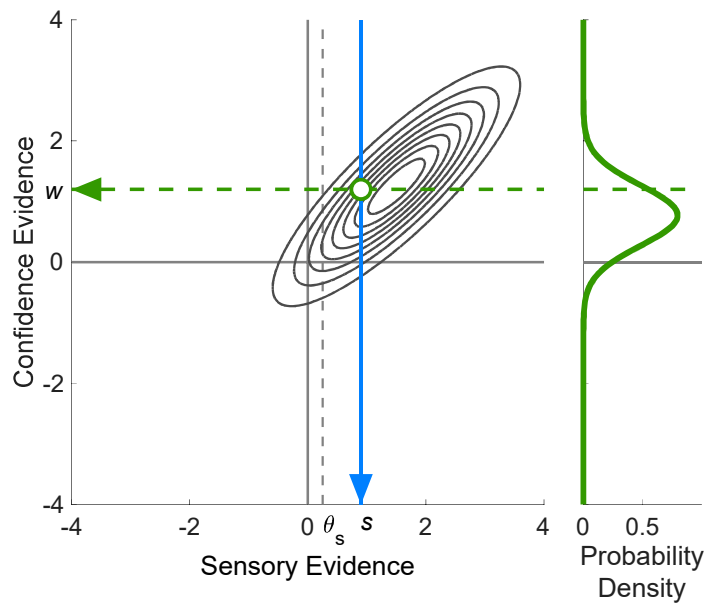
434 Taking into account all the possible deviations from the ideal confidence observer, the confidence  
 435 evidence is following Equation 9 above. This evidence is normally distributed with mean

$$436 \quad E[w] = (\mu_s - \theta_s - \theta_c) \beta / \sigma_s . \quad (10)$$

437 In addition, we note that confidence noise is independent of sensory evidence. This allows us to  
 438 characterize the variance of the distribution of confidence evidence as

$$439 \quad \text{var}[w] = (1 - \alpha)^2 \beta^2 + \sigma_c^2 . \quad (11)$$

440



441

442 Figure 7. Joint distribution of sensory and confidence evidence. On each trial, the  
 443 participant has access to one sensory sample  $s$  (blue arrow) and one confidence  
 444 sample  $w$  (green arrow) of the joint distribution  $H(s, w)$ . The blue distribution shown in  
 445 Figure 5 is the marginal distribution of the sensory evidence. The green distribution in  
 446 the right-hand panel is the distribution of confidence evidence for the particular  
 447 sensory sample  $s = 0.9$  (it is the cross-section of the joint distribution along the blue  
 448 line). The mean of this distribution is  $Q(s; \mu_A, \sigma_s)$  (see below, Equation 17), and its  
 449 spread is the confidence noise  $\sigma_c$ . The strength of the confidence evidence on that  
 450 particular trial is given by the magnitude of the sample  $w$  (distance away from zero).

451

452 Because both the sensory and confidence evidence are normally distributed, their joint distribution  
453  $H(s, w)$  is a bivariate normal distribution. An example of this joint distribution is shown in Figure 7.

454 The mean of the joint distribution  $H(s, w)$  is obtained from the mean of the sensory evidence and  
455 the mean of the confidence evidence (see Equation 10)

456 
$$E[ H(s, w) ] = [\mu_s, (\mu_s - \theta_s - \theta_c) \beta / \sigma_s] . \quad (12)$$

457 The covariance between  $s$  and  $w$  is obtained from Equation 9

458 
$$\text{cov}(s, w) = (1 - \alpha) \beta \sigma_s , \quad (13)$$

459 so that the covariance matrix  $K$  of the joint distribution  $H$  is

460 
$$K = \text{cov}[ H(s, w) ] = \begin{bmatrix} \sigma_s^2 & (1 - \alpha) \beta \sigma_s \\ (1 - \alpha) \beta \sigma_s & (1 - \alpha)^2 \beta^2 + \sigma_c^2 \end{bmatrix} . \quad (14)$$

461 It is worth noting the special case of the ideal confidence observer. In this case,  $\alpha = 0$ ,  $\beta = 1$ ,  $\sigma_c =$   
462  $0$ , and the covariance matrix reduces to

463 
$$K_{\text{ideal}} = \begin{bmatrix} \sigma_s^2 & \sigma_s \\ \sigma_s & 1 \end{bmatrix} . \quad (15)$$

464 The determinant of this covariance matrix is zero, indicating that there is a direct mapping between  
465 sensory evidence and confidence evidence: this is expected since without confidence noise,  
466 confidence and sensory evidence are perfectly correlated.

467 One other special case of interest is the super-ideal confidence observer ( $\alpha = 1$ ) corrupted with  
468 some confidence noise, where the covariance matrix is

469 
$$K_{\text{noisy\_super\_ideal}} = \begin{bmatrix} \sigma_s^2 & 0 \\ 0 & \sigma_c^2 \end{bmatrix} . \quad (16)$$

470 This covariance matrix is now diagonal, indicating that confidence and sensory evidences are  
471 independent. Here, the joint distribution  $H$  has its main axes oriented along the sensory and  
472 confidence evidence axes. In other words, for a noisy super-ideal confidence observer, confidence  
473 evidence depends only on the stimulus strength and is independent from the sensory evidence for  
474 the current trial.

475 6.b. Confidence Evidence Conditional on Sensory Evidence

476 On any perceptual trial of a confidence pair, the observer first gets some sensory evidence,  
 477 performs a perceptual decision based on this sensory evidence, and then estimates the confidence  
 478 that this decision is self-consistent. Therefore, we need to estimate the distribution of confidence  
 479 evidence for one particular value of sensory evidence  $s$  (Figure 7, right-hand panel). This  
 480 distribution of confidence evidence is  $P(w | s)$  and corresponds to a section of the joint distribution  
 481  $H(s, w)$ . This conditional distribution is normally distributed, and its mean (that we denote  
 482  $Q(s; \mu_s, \sigma_s)$  for later use) and variance can be inferred from the mean and the covariance matrix of  
 483 the joint distribution  $H$  (Equations 12 and 14)

$$484 \quad \begin{cases} E[w | s] = Q(s; \mu_s, \sigma_s) = [s + (\mu_s - s) \alpha - \theta_s - \theta_c] \beta / \sigma_s \\ \text{var}[w | s] = \sigma_c^2 \end{cases} \quad (17)$$

485 As expected, we see that the variance of the confidence evidence, once the sensory evidence is  
 486 known, is just the variance of the confidence noise. The mean is a biased and scaled version of  
 487 the sensory evidence  $s$ . It is biased towards the representation of the original stimulus  $\mu_s$  when the  
 488 parameter  $\alpha$  is larger than zero, i.e. when the human confidence observer is behaving a bit like the  
 489 super-ideal confidence observer. The scaling involves the parameter  $\beta$  that is responsible for a  
 490 proper calibration of confidence judgments, such that  $\beta > 1$  corresponds to over-confidence.

491 6.c. Pseudo-Perceptual Decision

492 The confidence evidence is the basis to judge whether the perceptual decision is self-consistent.  
 493 One might be tempted to just use the absolute value of confidence evidence for this judgment,  
 494 where larger absolute values reflect better chances to be self-consistent. However, this choice  
 495 would disregard the actual perceptual decisions that were taken. Critically, to decide whether the  
 496 perceptual decision is self-consistent, we need to evaluate whether the confidence evidence is  
 497 consistent with the perceptual decision. For this purpose, we introduce the *pseudo perceptual*  
 498 *decision*  $D'$  that corresponds to the perceptual decision that would have been taken if the  
 499 confidence evidence was used instead of the sensory evidence. By similarity to the definition of  
 500 perceptual decisions in Equation 2 above, the pseudo perceptual decision is thus defined as

$$501 \quad \begin{cases} D' = R & \text{if } w > 0, \\ D' = L & \text{otherwise} \end{cases} \quad (18)$$

502 When the pseudo perceptual decision  $D'$  is distinct from the perceptual decision  $D$ , this can be  
 503 taken as an alert signal that the perceptual decision might be invalid. Therefore, we can define a

504 new variable that reflects the diminished trust that the perceptual decision was valid when  $D'$  is  
505 distinct from  $D$ . We define the *signed confidence evidence* as

$$506 \quad w' = \begin{cases} |w| & \text{if } D = D', \\ -|w| & \text{otherwise} \end{cases} . \quad (19)$$

507 This signed confidence evidence is useful to estimate the probability that the perceptual decision is  
508 self-consistent given the current confidence evidence and perceptual decision,  
509  $P(\text{self-consistent} \mid w, D)$ . Computing this probability is complex because it rests on the knowledge  
510 of all the parameters in our model. Whereas prior work has assumed that observers would be able  
511 to use this knowledge (Fleming & Daw, 2017), here we propose instead that human observers only  
512 have access to the current level of confidence evidence and what they decided perceptually.  
513 Therefore, we propose that the observer is computing the *confidence probability* defined as

$$514 \quad P(\text{confident} \mid w, D) = \Phi(w') , \quad (20)$$

515 where  $\Phi$  is the cumulative of the standard normal distribution. In Appendix A, we show that the  
516 confidence probability is a reasonable proxy for the probability of being self-confident given the  
517 current confidence evidence and perceptual decision.

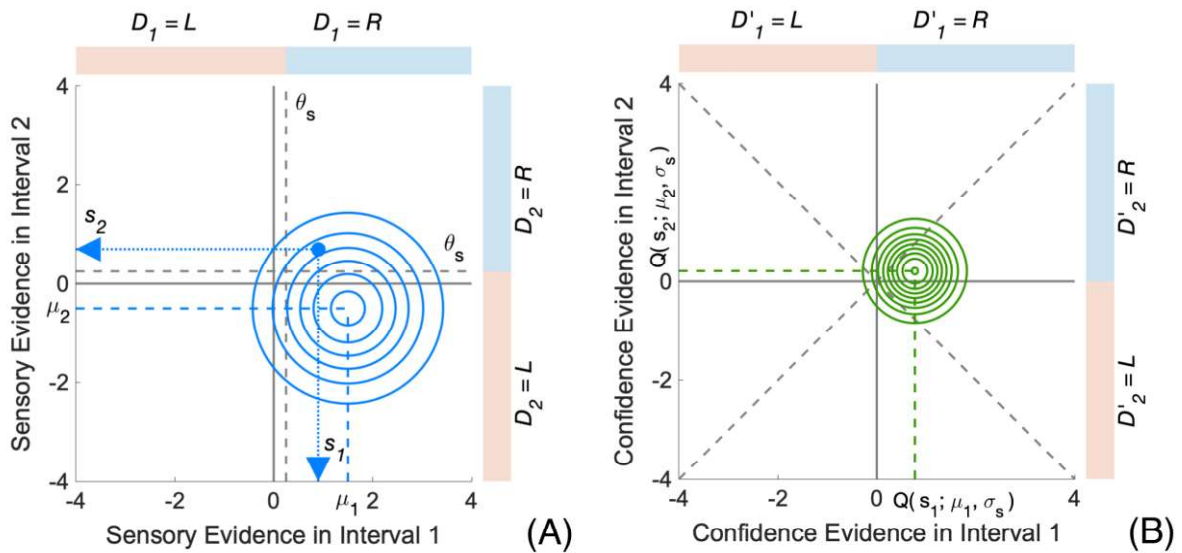
## 518 **7. Comparing Confidence Across Two Perceptual Decisions**

519 In the confidence forced-choice paradigm, two intervals are presented to the observer who has to  
520 choose the one for which she feels more confident that her perceptual decision was self-  
521 consistent. Therefore, we need to compare confidence across the two perceptual decisions of a  
522 confidence pair.

### 523 **7.a. Joint Sensory Evidence and Joint Confidence Evidence in a Confidence Pair**

524 Typically, the stimuli presented in the two intervals are independent from each other, so that we  
525 can assume that the sensory evidence in the two intervals is uncorrelated. Likewise, we assume  
526 that the confidence evidence in the two intervals is also uncorrelated. It is convenient to represent  
527 sensory and confidence evidence across the two intervals as joint probability distributions (Figure  
528 8).

529



530

531 Figure 8. Joint distributions of sensory and confidence evidence across the two  
 532 intervals of a confidence pair. (A) Joint distribution for the sensory evidence  $F(s_1, s_2)$ .  
 533 In this example, stimulus  $A$  is presented in interval 1 ( $\mu_1 = \mu_A$ ) and stimulus  $B$  is  
 534 presented in interval 2 ( $\mu_2 = \mu_B$ ), and are associated with the same level of sensory  
 535 noise ( $\sigma_1^2 = \sigma_2^2 = \sigma_s^2$ ). The joint distribution of the sensory evidence is shown as a  
 536 contour plot in blue. A sample of this joint distribution is shown as a blue dot that has  
 537 coordinates  $s_1$  for interval 1 and  $s_2$  for interval 2. The perceptual decisions  $D_1$  and  $D_2$   
 538 associated with this sample are both in favour of response  $R$ . (B) Joint distribution for  
 539 the confidence evidence conditional on sensory evidence  $G(w_1, w_2 | s_1, s_2)$ . Because  
 540 the perceptual decisions were  $R$  for both intervals, the joint confidence distribution is  
 541 likely to have its centre in the upper-right quadrant (contour plot in green). The pseudo  
 542 perceptual decisions  $D'_1$  and  $D'_2$  are shown for the confidence evidence space.

543

544 The joint distribution  $F(s_1, s_2)$  for the sensory evidence across the two intervals is a bivariate  
 545 normal distribution (Figure 8A) with mean and covariance

$$546 \begin{cases} E[F(s_1, s_2)] = [\mu_1, \mu_2] \\ \text{cov}[F(s_1, s_2)] = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \end{cases} \quad (21)$$

547 The joint distribution  $G(w_1, w_2 | s_1, s_2)$  is the confidence evidence conditional on the sensory  
 548 evidence across the two intervals (Figure 8B). It is a bivariate normal distribution with mean and  
 549 covariance matrix



$$\begin{cases} E[ G(w_1, w_2 | s_1, s_2) ] = [ Q(s_1; \mu_1, \sigma_1), Q(s_2; \mu_2, \sigma_2) ] \\ \text{cov}[ G(w_1, w_2 | s_1, s_2) ] = \begin{bmatrix} \sigma_c^2 & 0 \\ 0 & \sigma_c^2 \end{bmatrix} \end{cases} \quad (22)$$

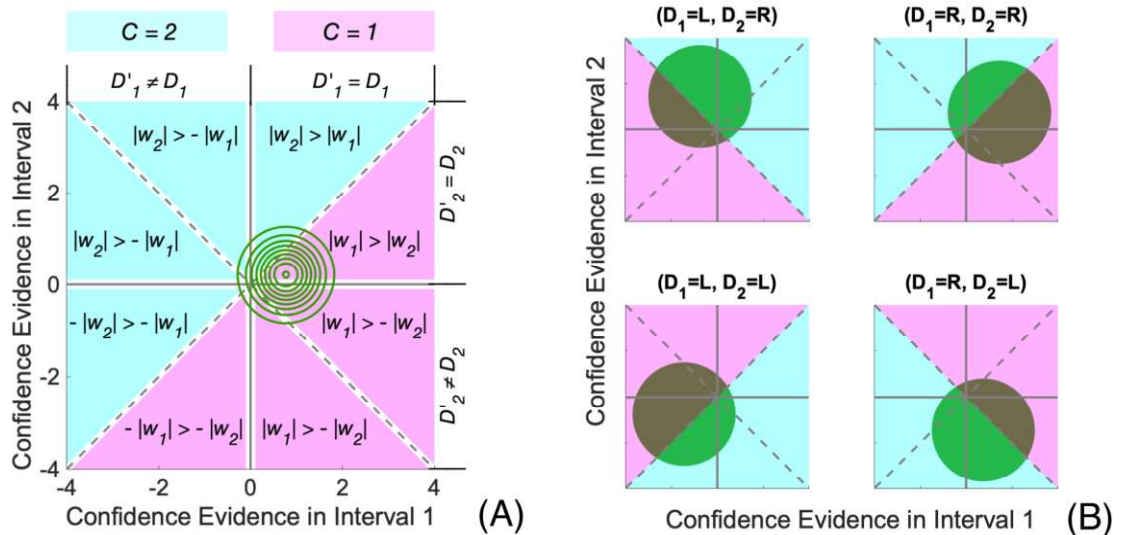
551 where the off-diagonal elements of the covariance matrix are zero because confidence evidence  
552 was assumed to be uncorrelated across intervals. The mean values are computed from Equation  
553 17.

#### 554 7.b. Confidence Decision Rule

555 The final step in choosing the interval in the confidence forced-choice paradigm is to decide on a  
556 *confidence decision rule*. This decision rule uses the confidence evidence in both intervals to  
557 select the interval the observer believes her perceptual decision is more self-consistent than the  
558 other. To take into account the perceptual decision in the confidence judgment, we rely on the  
559 signed confidence judgment  $w'$  described above (Equation 19). We define the choice of the  
560 confidence interval  $C$  between intervals 1 and 2 as follows

$$561 \quad C = \arg \max_{i \in \{1,2\}} (w'_i) \quad (23)$$

562 According to this Equation, the confidence choice will be the interval for which the confidence  
563 evidence is the largest in magnitude, except if there is a mismatch between  $D$  and  $D'$ , in which  
564 case the confidence choice will be the other interval. The impact of the inconsistency between  $D$   
565 and  $D'$  is illustrated in Figure 9A. This figure is reproduced from the previous example where the  
566 perceptual decisions were R in both intervals (Figure 8). Following Equation 23, interval 1 will be  
567 chosen if the confidence evidence lies in the contiguous half space in the lower-right. Applying the  
568 confidence decision rule to the other three scenarios of the perceptual decisions in intervals 1 and  
569 2 also leads to contiguous half-spaces that are consistent with a confidence choice in favour of one  
570 interval (Figure 9B).



571

572 Figure 9. Confidence decision rule. (A) Joint distribution for the confidence evidence  
 573 conditional on sensory evidence  $G(w_1, w_2 | s_1, s_2)$  when  $s_1$  and  $s_2$  are both consistent  
 574 with percept R. This plot is a replica of Figure 8B where eight different sectors are  
 575 identified from the comparison of the signed confidence evidence across the two  
 576 intervals. Sectors that lead to choosing interval 1 as more confident are shown in  
 577 purple ( $C = 1$ ), and those favouring interval 2 in cyan ( $C = 2$ ). Confident choices in  
 578 favour of interval 1 lie in a contiguous half-space located in the lower-right of the  
 579 confidence evidence space. (B) Confidence choices for each of the four possible  
 580 combinations of perceptual decisions across the two intervals. Labels of each panel  
 581 correspond to the perceptual decisions in each interval (e.g. “ $(D_1 = L, D_2 = R)$ ”  
 582 indicates that response category L was chosen in interval 1 and R in interval 2). The  
 583 scenario “ $(D_1 = R, D_2 = R)$ ” illustrated in part (A) of the figure is shown in the upper-  
 584 right panel.

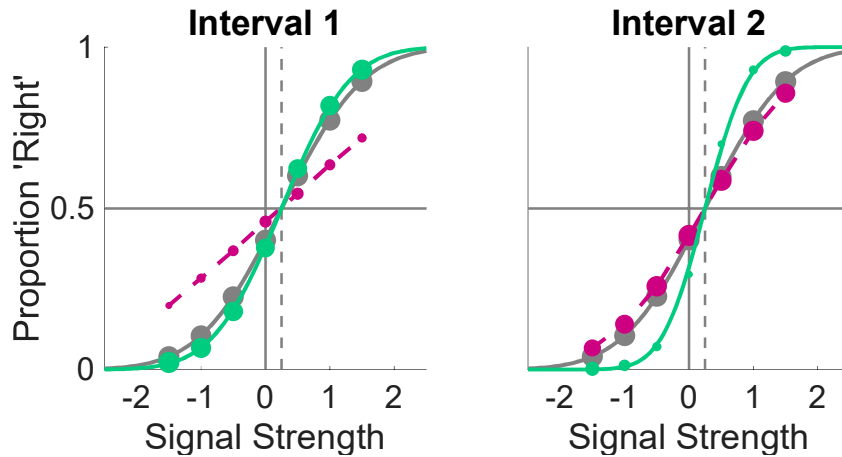
585

586 7.c. Interval Bias

587 We have to consider one last aspect of the confidence forced-choice paradigm. It is plausible that  
 588 participants will display some consistent bias in choosing the first or the second interval in all the  
 589 confidence trials. This type of interval bias has been found to be significant in some individuals,  
 590 and when it was present, it was relatively stable within individuals (de Gardelle & Mamassian,  
 591 2015). If we denote by  $\gamma$  the bias in favour of interval 1, then we can rewrite Equation 23 as follows

592 
$$\begin{cases} C = 1 & \text{if } w'_1 - w'_2 + \gamma > 0 \\ C = 2 & \text{otherwise} \end{cases} \quad (24)$$

593 When there is a bias to choose interval 1 over interval 2 ( $\gamma > 0$ ), interval 1 might be preferred over  
 594 interval 2 even when the perceptual decision in interval 2 was better than the one in interval 1. This  
 595 leads to worse discriminability of chosen decisions in interval 1 as compared to interval 2 (Figure  
 596 10).



598 Figure 10. Effect of interval bias on psychometric function. In these simulations, there  
 599 was a bias for the first interval ( $\gamma = 1.0$ ). The other parameters are listed in Table 2.  
 600 Plotting conventions are those of Figure 6.

601

602 The new division of confidence evidence space where intervals 1 and 2 are chosen should take  
 603 into account this interval bias (Appendix B).

## 604 8. Integrated Model for a Confidence Pair

605 So far, we have considered what is happening on a single confidence pair. In order to make  
 606 predictions from our model, we need to integrate all possible samples with their respective  
 607 distributions. This is equivalent to simulating our model with an infinite number of trials.

608 We start with the joint distribution  $G(w_1, w_2 | s_1, s_2)$  of confidence evidence conditional on the  
 609 sensory evidence across the two intervals. Equation 22 provides the mean and covariance of this  
 610 joint distribution. Following the confidence decision rule, the probability of choosing interval 1 as  
 611 more confident can be evaluated by integrating over the relevant part of the confidence space,  
 612 which depends on the perceptual decisions  $(D_1, D_2)$  (see Figure 9 and Appendix B). We need to  
 613 consider separately the four cases corresponding to the 2 by 2 possible perceptual decisions

614  $(D_1, D_2)$ . As detailed above, when there is no interval bias ( $\gamma = 0$ ), this space is simply a half-space  
 615 above or below one of the two diagonals (see Figure 9B). For instance, if both perceptual  
 616 decisions are 'R' (top-right panel in Figure 9B), we have

$$617 \quad P(C = 1 \mid s_1, s_2, D_1 = R, D_2 = R) = \int_{-\infty}^{+\infty} \int_{-\infty}^x G(x, y \mid s_1, s_2) dy dx \quad . \quad (25)$$

618 Obviously, the probability of choosing interval 2 as more confident is 1 minus this probability. With  
 619 a change of variables that rotates the confidence space by  $\pi/4$  counter-clockwise, the double  
 620 integral in Equation 25 can be reduced to a single integral

$$621 \quad \begin{aligned} P(C = 1 \mid s_1, s_2, D_1 = 'R', D_2 = 'R') \\ &= \int_{-\infty}^0 \varphi(v; (-Q(s_1; \mu_1, \sigma_1) + Q(s_2; \mu_2, \sigma_2))/\sqrt{2}, \sigma_c^2) dv \quad , \quad (26) \\ &= \Phi((Q(s_1; \mu_1, \sigma_1) - Q(s_2; \mu_2, \sigma_2))/(\sqrt{2} \sigma_c)) \end{aligned}$$

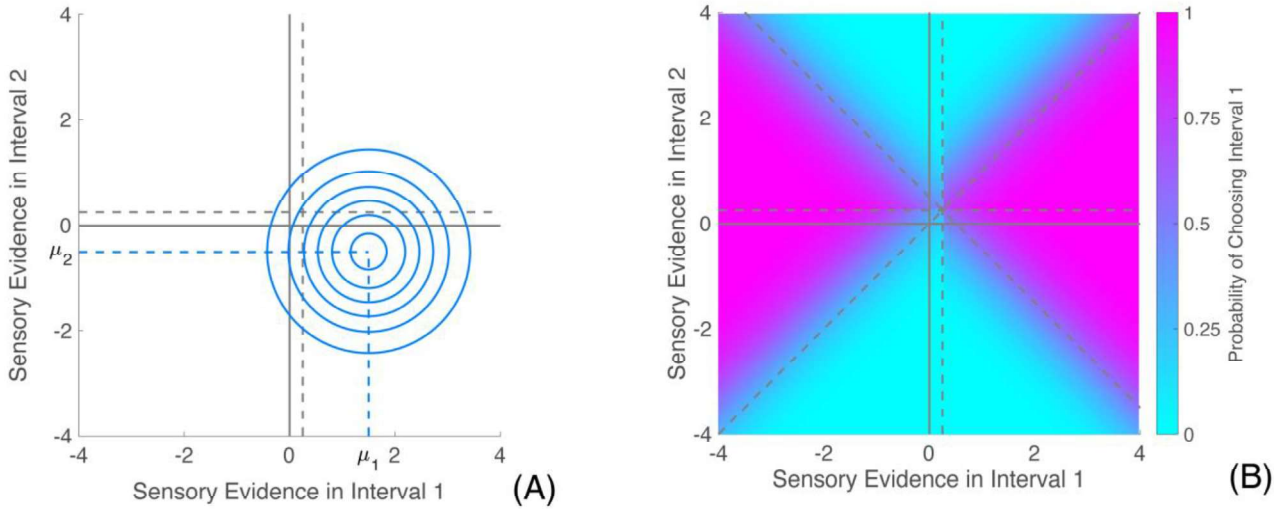
622 where  $\varphi(x; \mu, \sigma^2)$  is the probability distribution function of the normal distribution with mean  $\mu$  and  
 623 variance  $\sigma^2$ , and  $\Phi$  is again the cumulative distribution function of the standard normal distribution.  
 624 We can proceed similarly, for the three other cases to cover all possible pairs of perceptual  
 625 decisions in intervals 1 and 2,

$$626 \quad P(C = 1 \mid s_1, s_2, D_1, D_2) = \begin{cases} \Phi((Q(s_1; \mu_1, \sigma_1) - Q(s_2; \mu_2, \sigma_2))/(\sqrt{2} \sigma_c)) & \text{if } D_1 = 'R' \text{ \& } D_2 = 'R' \\ \Phi((Q(s_1; \mu_1, \sigma_1) + Q(s_2; \mu_2, \sigma_2))/(\sqrt{2} \sigma_c)) & \text{if } D_1 = 'R' \text{ \& } D_2 = 'L' \\ \Phi((-Q(s_1; \mu_1, \sigma_1) - Q(s_2; \mu_2, \sigma_2))/(\sqrt{2} \sigma_c)) & \text{if } D_1 = 'L' \text{ \& } D_2 = 'R' \\ \Phi((-Q(s_1; \mu_1, \sigma_1) + Q(s_2; \mu_2, \sigma_2))/(\sqrt{2} \sigma_c)) & \text{if } D_1 = 'L' \text{ \& } D_2 = 'L' \end{cases} \quad . \quad (27)$$

627 When there is an interval bias ( $\gamma \neq 0$ ), these conditional probabilities are still cumulative normal  
 628 functions, but over a larger or smaller domain (see Appendix B).

629 When we consider all the possible pairs of sensory evidence presented in the two intervals, we see  
 630 that the sensory criteria divide the sensory space into four quadrants (see again Equation 2).  
 631 Applying Equations 27 to the relevant quadrants produces the *confidence choice map* shown in  
 632 Figure 11B.

633



634

635 Figure 11. Joint distribution of sensory evidence and confidence choice map. (A) Joint  
 636 distribution of sensory evidence (replica of Figure 8A reproduced here for  
 637 convenience). (B) Confidence choice map. The probability of choosing interval 1 as  
 638 more confident is plotted for each pair of sensory evidence values in intervals 1 and 2.  
 639 Parameters for this example are listed in Table 2.

640

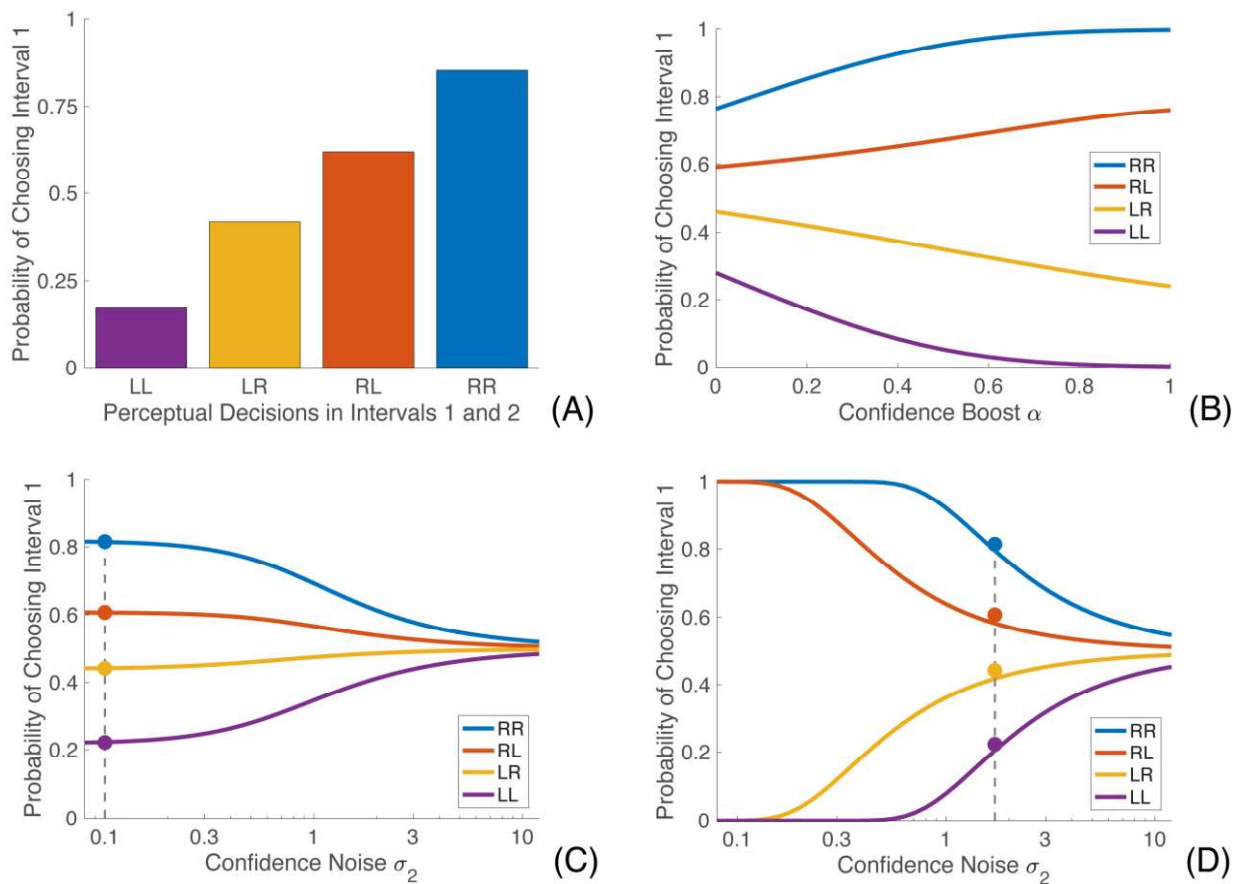
641 The final step to compute the integrated model is to combine the probability of getting a particular  
 642 pair of sensory evidence values  $(s_1, s_2)$  with its associated probability of choosing interval 1 as  
 643 more confident. The former is the joint distribution of sensory evidences across the two intervals  
 644 (Figure 11A) and the latter is the confidence choice map (Figure 11B). In layman's terms, we need  
 645 to multiply point by point Figure 11A with Figure 11B, and then integrate over the whole space.

646 In formal terms, the probability of choosing interval 1 as more confident is

647 
$$P(C = 1 | D_1, D_2) = \frac{\iint_{\Omega} P(C = 1 | s_1, s_2, D_1, D_2) \cdot P(s_1, s_2) ds_1 ds_2}{\iint_{\Omega} P(s_1, s_2) ds_1 ds_2} , \quad (28)$$

648 where  $\Omega$  is the quadrant of the space of sensory evidence across the two intervals that is  
 649 compatible with the pair of perceptual decisions  $(D_1, D_2)$ . For instance, when  $(D_1, D_2) = (R, R)$ ,  
 650  $\Omega = [\theta_s, +\infty) \times [\theta_s, +\infty)$ . We can easily compute a numerical approximation for this equation. The  
 651 result for the different perceptual decisions forms a quadruplet of probabilities as shown in Figure  
 652 12A.

653



654

655 Figure 12. Interval choice probabilities. (A) Quadruplet of confidence choice  
 656 probabilities for a particular pair of stimuli in the two intervals. The probability of  
 657 choosing interval 1 as more confidence is plotted for each pair of perceptual decisions  
 658 in intervals 1 and 2. Labels for the bars correspond to the perceptual decisions in each  
 659 interval (e.g. “LR” indicates that response category L was chosen in interval 1 and R in  
 660 interval 2). (B) Effect of confidence boost on interval choice probability. (C) Effect of  
 661 confidence noise on interval choice probability when the confidence boost is  $\alpha = 0$ . (D)  
 662 Effect of confidence noise on interval choice probability when the confidence boost is  
 663  $\alpha = 1$ . The four coloured dots in panels (C) and (D) have the same set of four values  
 664 of interval choice probabilities, therefore the corresponding pairs of confidence boost  
 665 and confidence noise are confidence metamers. All parameters, other than the  
 666 confidence boost in panels (B), (C) and (D), and the confidence noise in panels (C)  
 667 and (D), are listed in Table 2.

668

## 669 9. Confidence Metamers and Confidence Efficiency

### 670 9.a. Confidence Metamers

671 It is instructive to look at the effects of the two main parameters of the model, namely the  
672 confidence boost and the confidence noise while keeping the other parameters of the model  
673 constant (see Appendix C). Figure 12B illustrates how increasing confidence boost makes the  
674 probability of choosing interval 1 deviate from chance level (0.5), for each pair of perceptual  
675 decisions. Whether each of these probabilities tends towards 0 or 1 depends on the sign of  
676  $|\mu_1 - \theta_s| - |\mu_2 - \theta_s|$  (Appendix C).

677 Figures 12C and 12D illustrate the effect of confidence noise. As expected, increasing confidence  
678 noise makes confidence choices converge towards chance level. This convergence to chance  
679 level can be observed both when the confidence boost is small (Figure 12C) and large (Figure  
680 12D).

681 Comparing Figures 12C and 12D, we can see that confidence boost and confidence noise have  
682 opposite effects on interval choice probability. In other words, different pairs of confidence boost  
683 and confidence noise trade off and can produce similar outcomes in terms of confidence choice  
684 probabilities. One such example is shown with dashed lines in Figures 12C and 12D. These lines  
685 indicate that for an arbitrary choice of confidence boost and confidence noise ( $\alpha = 0$ ,  $\sigma_c = 0.1$ ),  
686 one can find other pairs of confidence boost and confidence noise (for instance,  $\alpha = 1$ ,  $\sigma_c = 1.71$ )  
687 that give rise to similar quadruplets of choice probabilities. We call these configurations *confidence*  
688 *metamers*.

689 Confidence metamers are pairs of confidence boost and confidence noise that correspond to  
690 similar levels of confidence in that they generate very similar quadruplets of confidence choices for  
691 all pairs of perceptual decisions in a confidence forced-choice paradigm. Confidence metamers  
692 highlight the difficulty in separating out the contribution of confidence boost and confidence noise  
693 in confidence judgments. However, one benefit of this concept is that it will allow us to define a  
694 confidence efficiency that combines the contributions of confidence boost and confidence noise.

### 695 9.b. Confidence Efficiency

696 Given quadruplets of confidence choices, sets of confidence metamers are obtained by choosing  
697 the value of confidence boost and searching for the confidence noise that best approximates the  
698 confidence choices. Three examples of confidence metamer sets are shown in Figure 13A

699 depicting the trade-off between confidence boost and confidence noise. The set of confidence  
700 metamers corresponding to the ideal confidence observer (blue curve in Figure 13A) is particularly  
701 important because it divides the (confidence noise, confidence boost) space into two parts. On its  
702 right are all the confidence metamers that are worse than the ideal confidence observer (green  
703 shaded region in Figure 13A), and on its left, the ones that are better (red shaded region). We will  
704 come back to this distinction shortly, after defining confidence efficiency.

705 Confidence metamers that are better than the ideal confidence observer (e.g. the red curve in  
706 Figure 13A) are special because, for these metamers, there exists no confidence noise that can  
707 lead to an equivalent confidence performance when the confidence boost is zero. Note however  
708 that all confidence metamer traces do cross the top horizontal line corresponding to the maximal  
709 confidence boost ( $\alpha = 1$ ; horizontal dashed line in Figure 13A). This property allows us to define  
710 the *equivalent confidence noise*  $\tau$  which is the confidence noise of the confidence metamer that  
711 corresponds to ( $\alpha = 1$ ). These equivalent confidence noises are shown as dots at the top of Figure  
712 13A. The blue dot is the equivalent confidence noise  $\tau_{\text{ideal}}$  for the ideal observer.

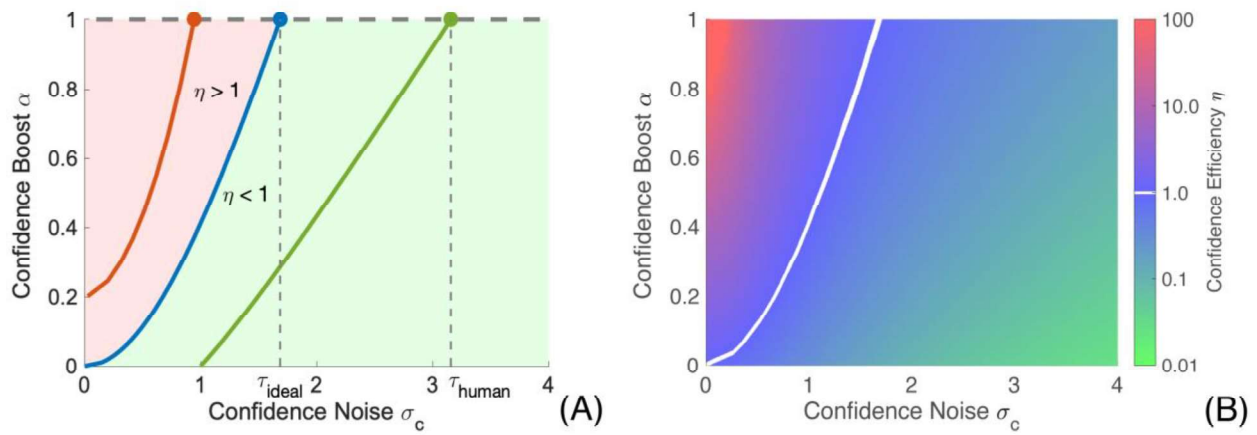
713 The equivalent confidence noise can help us summarize the sensitivity of the confidence  
714 judgments for a given set of confidence metamers, for instance the metamers shown in green in  
715 Figure 13A. We call this summary the *confidence efficiency*  $\eta$  that we define from the inverse of the  
716 equivalent confidence noise variance

$$717 \quad \eta = \tau_{\text{ideal}}^2 / \tau_{\text{human}}^2 . \quad (29)$$

718 In this definition, we have normalized the equivalent confidence noise of the human observer by  
719 that of the ideal confidence observer, so that the confidence efficiency is exactly 1 for the ideal  
720 confidence observer. The ratio of equivalent confidence noises is squared to make confidence  
721 efficiency analogous with the definition of efficiency for perceptual decisions (e.g. Kersten &  
722 Mamassian, 2009). Coming back to the two regions of Figure 13A defined by the ideal confidence  
723 observer, all confidence metamers to the right of the curve traced by the ideal confidence observer  
724 have a confidence efficiency smaller than 1, and those to its left have a confidence efficiency  
725 greater than 1.

726





727

728

729

730

731

732

733

734

735

736

737

738

739

740

741

742

743

744

745

Figure 13. Confidence metamers and confidence efficiency. (A) Construction of the equivalent confidence noise. Each of the three coloured curves shows confidence metamers, namely the pairs of confidence noise and confidence boost that produce similar quadruplets of choice probabilities across all four possible perceptual decisions of a confidence pair. The blue curve corresponds to the ideal confidence observer ( $\alpha = 0, \sigma_c = 0$ ). It intersects the line of maximal confidence boost ( $\alpha = 1$ ; horizontal dashed line at the top) at a point called the equivalent confidence noise for the ideal confidence observer ( $\tau_{\text{ideal}}$ ). For each confidence metamer, we can similarly find the equivalent confidence noise (e.g. the value  $\tau_{\text{human}}$  for the green curve that corresponds to a noisy ideal confidence observer ( $\alpha = 0, \sigma_c = 1$ )). (B) Confidence efficiency. The equivalent confidence noise can be used to compute the confidence efficiency (for the green confidence metamers in panel (A), the efficiency is  $\eta = 0.285$ ). By definition, confidence efficiency is 1 when both confidence boost and confidence noise are null. Confidence efficiency increases with confidence boost and decreases with confidence noise. Any pair of confidence noise and confidence boost that are to the right and below of the blue curve in panel (A) have a confidence efficiency smaller than 1, and those to the left and above have a confidence efficiency greater than 1.

746

747

748

749

750

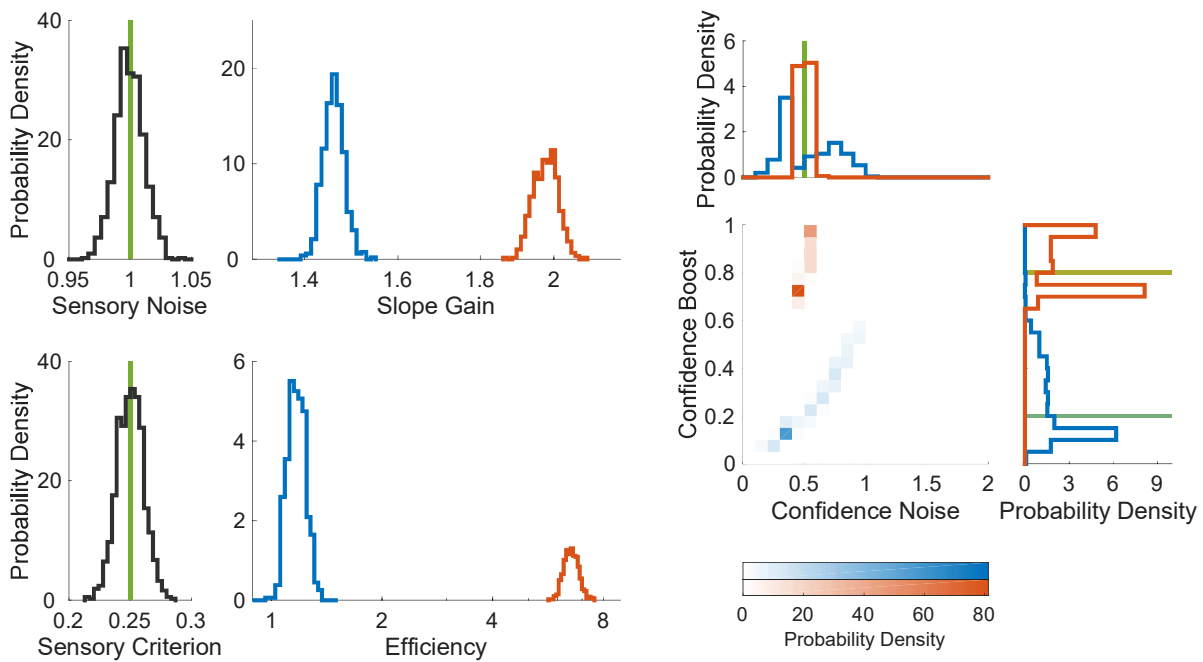
751

Using our definition of confidence efficiency, we can assign a confidence efficiency for each pair of confidence noise and confidence boost (Figure 13B). Confidence efficiency runs from zero (no metacognition, obtained when confidence noise is very large) to infinity (super-ideal confidence observer, obtained when confidence boost is 1 and there is no confidence noise). By definition, confidence efficiency is 1 for all the pairs of confidence noise and confidence boost that are confidence metamers of the ideal confidence observer.

752 **10. Full Model and Parameters Estimation**

753 When we introduced confidence metamers in the previous section, we discussed that confidence  
 754 boost and confidence noise were difficult to estimate simultaneously. There are however small  
 755 differences in the quadruplets of choice probabilities for different pairs of these parameters  
 756 (compare again Figure 12A with Figure 12B). In a real experiment that contains various stimulus  
 757 strengths that compete in confidence pairs, there will be a pair of confidence boost and confidence  
 758 noise that best explains all the choice probabilities.

759



760

761 Figure 14. Parameter recovery of the model. The distributions of parameters were  
 762 estimated from 500 simulated experiments. The estimated parameters were the  
 763 sensory noise  $\sigma_s$  and the sensory criterion  $\theta_s$  (first column), the gain in the slope of the  
 764 psychometric functions between chosen and unsorted trials and the confidence  
 765 efficiency (second column). The full confidence model also attempted to infer the  
 766 confidence boost  $\alpha$  and the confidence noise  $\sigma_c$  (right panel). Estimated confidence  
 767 boost and confidence noise are correlated, and this correlation creates confidence  
 768 metamers. The original parameter values that were used in the simulations are shown  
 769 as green lines. Two different confidence boosts were simulated,  $\alpha = 0.2$  in blue and  
 770  $\alpha = 0.8$  in orange. The other parameters are listed in Table 2.

771

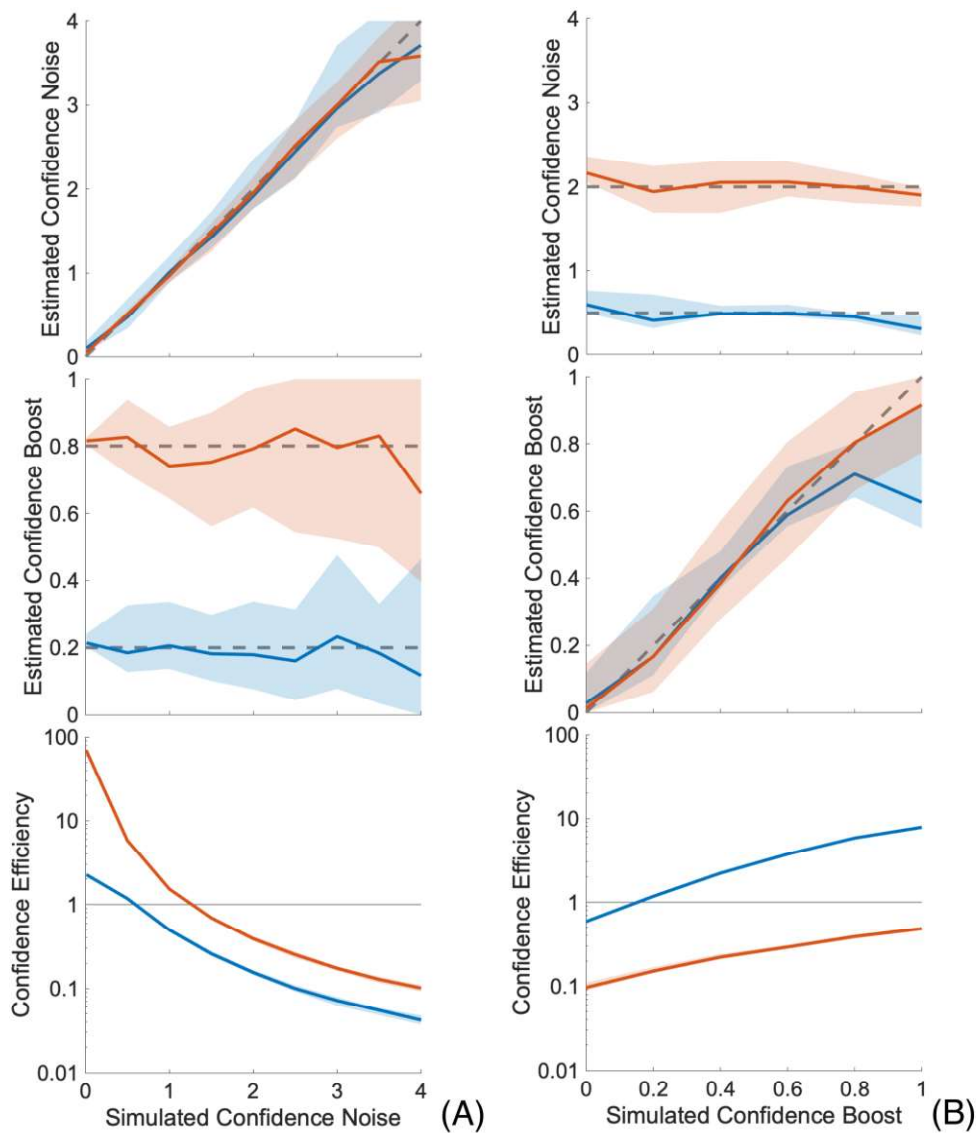
772 Assuming that the confidence pairs are independent from each other, we can obtain the set of best  
773 model parameters by summing the log likelihood of each confidence pair. An example of best fitted  
774 estimate is shown superimposed on the simulated data in Figure 3. In that figure, simulated  
775 parameters were  $\sigma_s = 1.0$ ,  $\sigma_c = 0.5$ ,  $\theta_s = 0.25$ , and  $\alpha = 0.2$ . Estimated parameters were  $\hat{\sigma}_s =$   
776  $0.999$ ,  $\hat{\sigma}_c = 0.326$ ,  $\hat{\theta}_s = 0.245$ ,  $\hat{\alpha} = 0.122$ , and  $\hat{\eta} = 0.789$  ( $\theta_c$ ,  $\beta$ , and  $\gamma$  were fixed to their default  
777 values). We see that estimated parameters are near their theoretical values, but there are small  
778 deviations.

779 To appreciate the faithfulness of our model parameters, we simulated 500 experiments with the  
780 same original parameters, and collected the distributions of the estimated parameters. Figure 14  
781 shows these distributions for two different values of confidence boost ( $\alpha = 0.2$  vs.  $\alpha = 0.8$ ). We  
782 observe that these two values of confidence boost can be distinguished since their distributions do  
783 not overlap. In addition, both the gain in the slope of the psychometric functions and the efficiency  
784 measures are able to distinguish these two conditions, since the distributions are clearly  
785 segregated (middle column of Figure 14).

786 The next figure shows simulations of the model with varying levels of confidence noise or varying  
787 levels of confidence boost (Figure 15). Critically, the estimated confidence noise follows very well  
788 the actual confidence noise for the two levels of confidence boost simulated (Figure 15A, top), and  
789 these levels of confidence boost are well-recovered independently of the confidence noise (Figure  
790 15A middle). The opposite holds when varying the confidence boost (Figure 15B). In short, both  
791 confidence noise and confidence boosts can be recovered very well.

792 In Appendix D, we present parameter recovery for the remaining parameters of the model. The  
793 confidence noise and boost parameters are quite stable for different values of sensory noise. This  
794 is not surprising since, in the model, confidence evidence is normalized by sensory sensitivity, so  
795 the confidence noise and boost parameters should not depend on sensory noise. The confidence  
796 noise and boost parameters are also quite stable for different values of sensory and confidence  
797 criteria, at least as long as these criteria are within reasonable limits of the range of the presented  
798 sensory stimuli. Importantly, the confidence noise and boost parameters are very stable for  
799 different values of biases in favour of responding either the first or second interval. In this latter  
800 case though, confidence efficiency decreases as the interval response bias increases, because  
801 favouring one interval over the other necessarily impairs the accuracy of choosing the interval that  
802 was more likely to be self-consistent. Finally, the confidence noise and boost parameters are better  
803 recovered as more confidence pairs are tested in an experiment. If the number of confidence pairs  
804 is less than about 1,000, the confidence noise and boost parameters are estimated to  
805 imprecisely, although the confidence efficiency remains a robust measure of meta-perception.

806



807

808

809

810

811

812

813

814

815

816

817

818

819

Figure 15. Model recovery for a range of confidence noise and confidence boost. (A) The plots show estimated parameters for two different values of confidence boost,  $\alpha = 0.2$  in blue and  $\alpha = 0.8$  in red. The estimated parameters are confidence noise (top panel), confidence boost (middle), and efficiency (bottom). (B) The plots show estimated parameters for two different values of confidence noise,  $\sigma_c = 0.5$  in blue and  $\sigma_c = 2.0$  in red. The thick lines are median estimated values across  $N = 40$  repeated simulations, and the shaded areas cover the 25<sup>th</sup> to the 75<sup>th</sup> interquartile range.

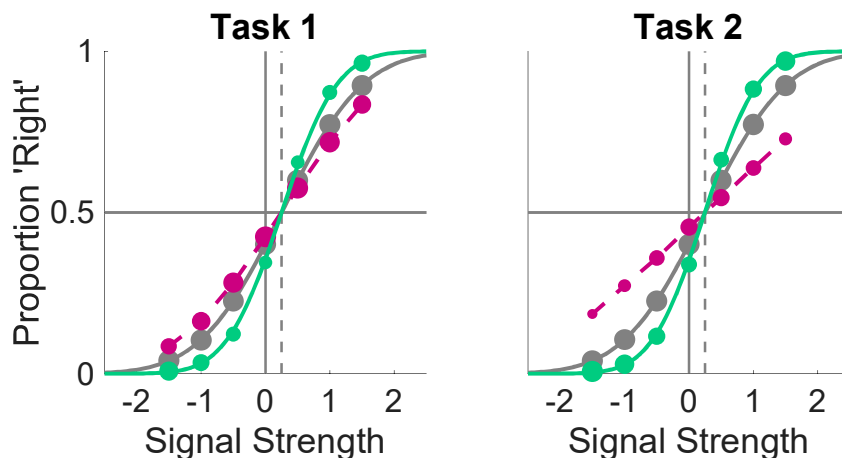
At this stage, we have not presented the model recovery for the last parameter of the model, the confidence bias  $\beta$ . This is because this scaling factor affects both intervals equally, so its effects cancel out in the confidence forced-choice paradigm (see section 5.c). In a sense, the confidence forced-choice paradigm was designed to be immune to possible confidence biases, so it was

820 expected that this bias would be difficult to estimate. However, there is one scenario where the  
821 confidence bias can be recovered, at least up to a scaling factor, and this is what we explore next.

## 822 11. Effects of Confidence Bias

823 So far, we have considered that participants were performing the same perceptual task in both  
824 intervals of a confidence pair. However, it is interesting to consider the condition where the  
825 participant is asked to perform different tasks across the two intervals. This condition allowed us to  
826 claim that confidence was computed in a common currency, rather than in some metric that is  
827 tightly constrained by the dimension along which the task is performed (de Gardelle & Mamassian,  
828 2014; de Gardelle, Le Corre, & Mamassian, 2016).

829



830

831 Figure 16. Effect of confidence bias on the psychometric functions. In these  
832 simulations, the first task was properly scaled ( $\beta = 1.0$ ) but the observer was over-  
833 confident in the second task ( $\beta = 2.0$ ). As a result, whenever task 1 is competing with  
834 task 2 in a confidence pair, confidence choice is biased in favour of task 2 (indicated  
835 by larger green dots for task 2 than for task 1). All parameters except  $\beta$  are identical  
836 across the two tasks and listed in Table 2. Plotting conventions are those of Figure 6.

837

838 A between-task confidence judgment also allows us to tackle an issue that we had to leave out  
839 when participants were performing the same task in both intervals of a confidence pair. This issue  
840 is whether participants are properly estimating their perceptual sensitivity in a task and correctly  
841 using this estimate to normalize their confidence evidence. In the model we described above, we  
842 assumed that this normalizing parameter  $\beta$  was indeed 1.0 (no confidence bias). If only one task is  
843 used, the effects of this parameter are invisible (Figure 6C), because the same scaling is applied to

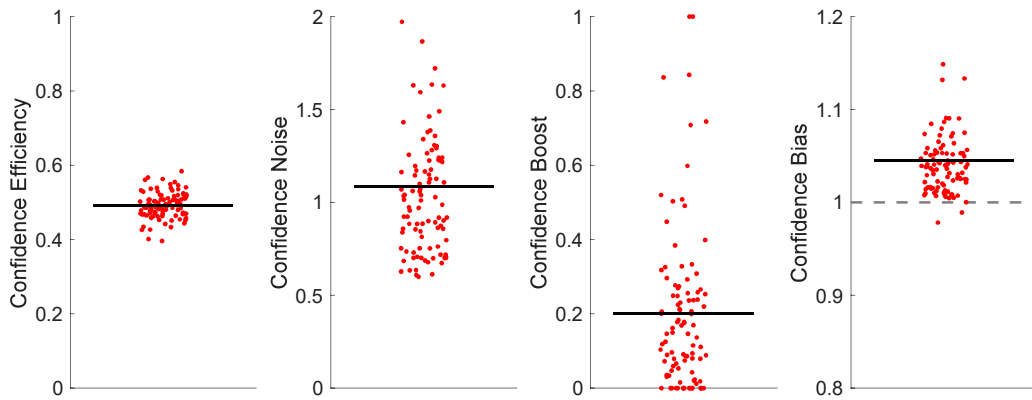
844 both confidence evidences of the two intervals. When two tasks are competing in the two intervals,  
845 there is a visible effect on the psychometric functions (Figure 16). When two tasks are run, we  
846 cannot estimate both corresponding  $\beta$  parameters, but we can estimate their ratios (see  
847 Appendix E). This allows us to estimate whether one task shows over- or under-confidence relative  
848 to the other task.

## 849 **12. Re-Analysis of De Gardelle & Mamassian (2015)**

850 So far, we have looked at the ability of the model to simulate a confidence forced-choice  
851 experiment, and the faithfulness of the recovered parameters. We now apply this framework to the  
852 re-analysis of one of our previous studies. We choose the study of confidence for motion direction  
853 discrimination that was published in de Gardelle & Mamassian (2015). In that experiment,  
854 observers had to discriminate the mean direction of motion above or below a reference for a  
855 stimulus composed of multiple random dot motion. The strength of the stimulus was manipulated  
856 by varying the mean motion direction, where larger mean motion directions away from the  
857 reference are easier stimuli. In addition, there were two stimulus uncertainty levels, represented by  
858 the different ranges of motion directions of the dots within a stimulus. Given that these ranges are  
859 very different, we can apply the analysis of confidence biases that we discussed in section 11,  
860 where the two tasks correspond here to the two stimulus uncertainty levels.

861 We present here parameter estimates based on the group data. This group data set corresponds  
862 to the data collected across all participants, after normalising each participant to her own sensory  
863 noise and criterion. The analysis thus assumes that there is single set of model parameters shared  
864 across all participants. In this sense, this analysis can be seen as complementary to the one  
865 presented in the original paper (de Gardelle & Mamassian, 2015), where individual differences  
866 were emphasized.

867 Parameter estimates for this experiment are shown in Figure 17. Confidence efficiency was about  
868 0.5, indicating that participants were clearly able to make meta-perceptual judgments (efficiency  
869 larger than 0) but less efficient than the ideal confidence observer (efficiency less than 1).  
870 Separating confidence efficiency into confidence noise and confidence boost, we found evidence  
871 that confidence in this task and for this stimulus was processed more in serial than in parallel to the  
872 perceptual decision (confidence boost closer to zero than to one). Confidence noise was estimated  
873 to be about 1 (this value does not have any unit and thus could potentially be compared to other  
874 confidence noise in other experiments). Finally, we also found a small but significant confidence  
875 bias, revealing an overconfidence for the high stimulus uncertainty relative to low stimulus  
876 uncertainty. In other words, on average, participants did not fully appreciate the effect of the  
877 stimulus noise on their sensory sensitivity.



878

879 Figure 17. Model parameter estimates in a real study. Individual dots are estimates  
 880 from 100 bootstrapped trials on the data collected over 15 observers. Data are from de  
 881 Gardelle & Mamassian (2015).

882

883 **13. Discussion**

884 In summary, we have presented here a generative model for the estimation of confidence in  
 885 perceptual decisions. Our model considers confidence to be the evaluation that one's perceptual  
 886 decision is self-consistent, thereby highlighting that confidence is about a decision, not about the  
 887 stimulus itself, its sensory uncertainty, contrast, duration or visibility. The self-consistency aspect of  
 888 the definition emphasizes that the perceiver evaluates her own percept, rather than whether her  
 889 percept is consistent with the true state of the world. Using this definition, we have proposed a  
 890 model of perceptual confidence where the perceptual decision follows classical Signal Detection  
 891 Theory (Green & Swets, 1966). We then assumed that confidence evidence scales with the  
 892 distance between sensory evidence and the sensory criterion, where the scaling factor is inversely  
 893 proportional to sensory noise. This confidence evidence is corrupted by confidence noise but can  
 894 benefit from some confidence boost that corresponds to the possibility that confidence may rely on  
 895 additional information compared to the sensory evidence. We identify three keys aspects by which  
 896 our approach goes beyond previous work.

897 First, we can theoretically differentiate between parallel and serial processing of confidence. To  
 898 obtain this result, we described the behavior of an ideal agent that uses the same information as  
 899 that used for the perceptual decision. This ideal confidence observer was contrasted to a super-  
 900 ideal agent that uses a novel and perfect estimation of the stimulus for the purpose of the  
 901 confidence judgment. Serial processing mimics the ideal confidence observer, albeit not optimally  
 902 (see also Bang et al., 2019), whereas parallel processing mimics the super-ideal confidence

903 observer. The fraction of ideal and super-ideal observers in the confidence judgments is  
904 represented by the confidence boost parameter in our model. To be precise, this parameter  
905 reflects the fraction of all information used in confidence processing that was not used for the  
906 perceptual decision (see also Barrett, Dienes & Seth, 2013; Maniscalco & Lau, 2016; Fleming &  
907 Daw, 2017). As such, it may aggregate information from multiple sources, including non-sensory  
908 information such as motor signals (see e.g. Fleming et al., 2015; Wokke et al., 2020), or  
909 fluctuations of attention (see e.g. Recht et al., 2019) or sensory information that was processed  
910 after the perceptual decision took place (Baranski & Petrusic, 1998; Pleskac & Busemeyer, 2010).  
911 Similarly, we should emphasize that the noise corrupting the confidence evidence, although  
912 quantified with a single parameter in our model, may aggregate multiple sources of inefficiencies,  
913 including noisy read-out of the perceptual evidence, but also influences from previous confidence  
914 judgments (Rahnev et al., 2015), or influences from other features that are not related to  
915 perceptual performance. Importantly, the confidence boost parameter was well recovered in our  
916 simulations that contained a large number of trials. We anticipate that the ability to distinguish  
917 between parallel and serial confidence processing will be an important asset of our model.

918 Second, we propose a measure of efficiency that is genuinely anchored to the metacognitive level  
919 of computation. Our efficiency measure is obtained by comparing human confidence performance  
920 to that of the ideal confidence observer. Along the way, we have defined confidence metamers that  
921 correspond to different observers who share the same confidence efficiency. Confidence  
922 metamers result from different trade-offs between two parameters of our model, confidence noise  
923 and confidence boost, and are hard to differentiate in an experiment that contains only a limited  
924 number of trials. Our definition of confidence efficiency deviates from previous ones. For instance,  
925 in the now popular *meta-d'* framework for analyzing confidence judgments (Maniscalco & Lau,  
926 2012), no generative model is specified for confidence judgments. Under that framework, *meta-d'*  
927 quantifies the sensitivity at the metacognitive level by estimating the first-order sensitivity that  
928 would be needed to observe the data if the metacognitive system were perfect. The *M-ratio*, that is  
929 the ratio of *meta-d'* over *d'*, has been put forward as a measure of efficiency, but although it makes  
930 some intuitive sense, it does not correspond to a clear process. Other theoretical approaches to  
931 metacognition have described potential generative models for confidence judgments (e.g. Pleskac  
932 & Busemeyer, 2010; Fleming & Daw, 2017; Sanders et al., 2016), but they did not offer an  
933 efficiency measure based on these models.

934 Third, our model can sometimes recover the confidence bias that corresponds to the mis-  
935 estimation of one's perceptual sensitivity. In our model, perceptual sensitivity is used to normalize  
936 confidence so that this latter can be compared across tasks and sensory modalities (de Gardelle &  
937 Mamassian, 2014). As a consequence, overconfidence corresponds here to an over-estimation of  
938 one's perceptual sensitivity. While the effects of confidence bias are invisible when one considers  
939 only one task, the ratio of confidence biases can be estimated when two tasks are compared.



940 Confidence comparison between two tasks is particularly easy within the confidence forced-choice  
941 paradigm. In this paradigm, a confidence choice is taken between two perceptual decisions. Using  
942 our modelling framework, we have described the probabilities with which one perceptual decision  
943 is associated with a larger confidence than the other decision, for different stimulus strengths and  
944 different commitments to perceptual decisions. Previous analyses of metacognitive abilities have  
945 had troubles to take into account varying difficulty levels. For instance, the classic measure of  
946 confidence resolution simply compares confidence in correct responses and errors, and ignores  
947 task difficulty. In the *meta-d'* approach, one major limitation is that it is designed to analyze data  
948 where perceptual sensitivity is constant across trials (only one stimulus strength is used in the  
949 experiment). Failure to meet this assumption leads to overestimations of metacognitive sensitivity  
950 (see e.g. Rahnev & Fleming, 2019), because participants could be using variations of performance  
951 that cannot be used in the *meta-d'* estimation procedure. Our method may allow researchers to  
952 overcome this obstacle.

953 Our model involves a number of parameters and assumptions, which deserve scrutiny. We argue  
954 however that most assumptions of our model are relatively standard and supported by empirical  
955 evidence. Besides, the parameters we have introduced all have a clear interpretation, and can be  
956 recovered quite well (see section 10 and Appendices D and E). The output of our model is a  
957 signed confidence evidence that approximates the probability that the perceptual decision is self-  
958 consistent. When applied to the confidence forced-choice paradigm, the decision rule for  
959 confidence is a simple comparison of the signed confidence evidence between two trials, and does  
960 not involve complex inference. In this respect, our approach appears less demanding than the  
961 *actor-critic* model of Fleming & Daw (2017) where confidence judgments require an inference  
962 based on the confidence evidence and the knowledge of the covariance between confidence  
963 evidence and sensory evidence. It is arguably unrealistic to assume that human participants have  
964 access to this latter knowledge, and it becomes computationally intense when multiple levels of  
965 difficulty are involved.

966 One aspect of our model that appears non trivial is the possibility that participants would use  
967 distinct decision criteria for the Type 1 response and for the Type 2 evaluation. This possibility was  
968 explicitly excluded in the *meta-d'* framework. Our framework allows for it, although we anticipate  
969 that a reduced model without this additional criterion should suffice in most case. However, this  
970 parameter might be interesting to researchers in some situations, where participants have to  
971 combine sensory and non-sensory information about a stimulus. The non-sensory information can  
972 be a probabilistic cue, as in many decision making studies (e.g. Locke et al., 2020), or an advice  
973 given by another observer, as for instance in Asch's conformity experiment (Asch 1956). Here, as  
974 they face a tradeoff between optimality and accuracy, participants might use a Type 1 criterion that  
975 takes into account all the cues to make their own decision, but a Type 2 criterion that only  
976 considers their own sensory information when evaluating their confidence. Future research, both

977 theoretical and empirical, may aim at understanding how metacognition unfolds in these situations  
978 of decision under influence.

979 To conclude, our effort has focused on specifying a formal generative model where confidence can  
980 be both corrupted and boosted relative to the sensory evidence, and the application of this model  
981 to the confidence forced choice paradigm. Obviously, this generative model could be used to  
982 derive confidence ratings on a scale, which are most commonly used in experiments. Doing so  
983 would require introducing additional parameters for the mapping between internal and reported  
984 confidence (Aitchison et al., 2015), which the confidence forced choice paradigm naturally avoids.  
985 One other direction for future work is to extend the present model to other perceptual tasks,  
986 including detection tasks (see e.g. García-Pérez et al., 2011). Finally, since the simultaneous  
987 estimation of all parameters in our model require a large amount of data, the development of a  
988 Bayesian hierarchical estimation would be important to be able to collect data across participants  
989 (Fleming, 2017). Ultimately, it will be interesting to compare the parameters of the generative  
990 model across tasks, sensory modalities, and participant populations.

## 991 **Acknowledgments**

992 This work was supported by grants ANR-10-BLAN-1910 “Visual Confidence” to PM, and ANR-18-  
993 CE28-0015 “VICONTE” to PM and VdG. Supplementary support came from ANR-17-EURE-0017.  
994 The authors would like to thank Tarryn Balsdon, Thibault Gajdos, Mike Landy, Shannon Locke,  
995 and Jérôme Sackur for their critical comments on an earlier draft of the manuscript.

996 Source code in Matlab for model fitting is available at: <https://github.com/mamassian/cfc>

997 **References**

- 998 Adler, W. T., & Ma, W. J. (2018). Limitations of Proposed Signatures of Bayesian Confidence.  
999 *Neural Computation*, 1–28.
- 1000 Aguilar-Lleyda, D., Lemarchand, M., & De Gardelle, V. (2020). Confidence as a priority signal.  
1001 *Psychological Science*, 31(9), 1084-1096.
- 1002 Aitchison, L., Bang, D., Bahrami, B., & Latham, P. E. (2015). Doubly Bayesian analysis of  
1003 confidence in perceptual decision-making. *PLoS Computational Biology*, 11(10), e1004519.  
1004 <http://doi.org/10.1371/journal.pcbi.1004519>
- 1005 Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a  
1006 unanimous majority. *Psychological Monographs: General and Applied*, 70(9), 1–70.
- 1007 Bang, J. W., Shekhar, M., & Rahnev, D. A. (2019). Sensory noise increases metacognitive  
1008 efficiency. *Journal of Experimental Psychology: General*, 148(3), 437–452.
- 1009 Baranski, J. V., & Petrusic, W. M. (1998). Probing the locus of confidence judgments: Experiments  
1010 on the time to determine confidence. *Journal of Experimental Psychology: Human Perception  
1011 and Performance*, 24(3), 929–945.
- 1012 Barlow, H. B. (1962). A method of determining the overall quantum efficiency of visual  
1013 discriminations. *The Journal of Physiology*, 160(1), 155–168.
- 1014 Barrett, A. B., Dienes, Z., & Seth, A. K. (2013). Measures of metacognition on signal-detection  
1015 theoretic models. *Psychological Methods*, 18(4), 535–552.
- 1016 Barthelmé, S., & Mamassian, P. (2009). Evaluation of objective uncertainty in the visual system.  
1017 *PLoS Computational Biology*, 5(9), 1–8. <http://doi.org/10.1371/journal.pcbi.1000504>
- 1018 Clarke, F. R., Birdsall, T. G., & Tanner, W. P., Jr. (1959). Two types of ROC curves and definitions  
1019 of parameters. *The Journal of the Acoustical Society of America*, 31, 629–630.
- 1020 de Gardelle, V., Le Corre, F., & Mamassian, P. (2016). Confidence as a common currency  
1021 between vision and audition. *PLoS ONE*, 11(1), e0147901.  
1022 <http://doi.org/10.1371/journal.pone.0147901>
- 1023 de Gardelle, V., & Mamassian, P. (2014). Does confidence use a common currency across two  
1024 visual tasks? *Psychological Science*, 25(6), 1286–1288.

- 1025 de Gardelle, V., & Mamassian, P. (2015). Weighting mean and variability during confidence  
1026 judgments. *PLoS ONE*, *10*(3), e0120870. <http://doi.org/10.1371/journal.pone.0120870>
- 1027 De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based  
1028 choice. *Nature neuroscience*, *16*(1), 105-110.
- 1029 Fleming, S. M. (2017). HMeta-d: hierarchical Bayesian estimation of metacognitive efficiency from  
1030 confidence ratings. *Neuroscience of Consciousness*, *3*(1:nix007), 1–14.  
1031 <http://doi.org/10.1093/nc/nix007>
- 1032 Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian  
1033 framework for metacognitive computation. *Psychological Review*, *124*(1), 91–114.
- 1034 Fleming, S. M., Dolan, R. J., & Frith, C. D. (2012). Metacognition: computation, biology and  
1035 function. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594),  
1036 1280–1286.
- 1037 Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human*  
1038 *Neuroscience*, *8*, 1–9. <http://doi.org/10.3389/fnhum.2014.00443>
- 1039 Fleming, S. M., Maniscalco, B., Ko, Y., Amendi, N., Ro, T., & Lau, H. C. (2015). Action-specific  
1040 disruption of perceptual confidence. *Psychological Science*, *26*(1), 89–98.
- 1041 Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal  
1042 detectability: Discrimination between correct and incorrect decisions. *Psychonomic Bulletin &*  
1043 *Review*, *10*(4), 843–876.
- 1044 García-Pérez, M. A., Alcalá-Quintana, R., Woods, R. L., & Peli, E. (2011). Psychometric functions  
1045 for detection and discrimination with and without flankers. *Attention, Perception &*  
1046 *Psychophysics*, *73*(3), 829–853.
- 1047 Geisler, W. S. (1989). Sequential ideal-observer analysis of visual discriminations. *Psychological*  
1048 *Review*, *96*(2), 267–314.
- 1049 Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. New York:  
1050 Wiley.
- 1051 Hainguerlot, M., Vergnaud, J. C., & de Gardelle, V. (2018). Metacognitive ability predicts learning  
1052 cue-stimulus associations in the absence of external feedback. *Scientific reports*, *8*(1), 1-8.
- 1053 Kersten, D., & Mamassian, P. (2009). Ideal observer theory. In L. R. Squire (Ed.), *Encyclopedia of*  
1054 *Neuroscience*, volume 5 (pp. 89-95). Oxford: Academic Press.

- 1055 Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review*,  
1056 119(1), 80–113.
- 1057 Locke, S. M., Gaffin-Cahn, E., Hosseinizaveh, N., Mamassian, P., & Landy, M. S. (2020). Priors  
1058 and payoffs in confidence judgments. *Attention, Perception & Psychophysics*, 77(2), 638–18.
- 1059 Mamassian, P. (2016). Visual confidence. *Annual Review of Vision Science*, 2(1), 459–481.
- 1060 Mamassian, P. (2020). Confidence forced-choice and other metaperceptual tasks. *Perception*,  
1061 49(6), 616–635.
- 1062 Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating  
1063 metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, 21(1), 422-  
1064 430.
- 1065 Maniscalco, B., & Lau, H. C. (2016). The signal processing architecture underlying subjective  
1066 reports of sensory awareness. *Neuroscience of Consciousness*, 2016(1):niw002, 1–17.  
1067 <http://doi.org/10.1093/nc/niw002>
- 1068 Meyniel, F., Sigman, M., & Mainen, Z. F. (2015). Confidence as Bayesian probability: From neural  
1069 origins to behavior. *Neuron*, 88(1), 78–92.
- 1070 Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*,  
1071 115(2), 502.
- 1072 Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings.  
1073 *Psychology of Learning and Motivation*, 26, 125–173.
- 1074 Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice,  
1075 decision time, and confidence. *Psychological Review*, 117(3), 864–901.
- 1076 Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: Distinct probabilistic  
1077 quantities for different goals. *Nature Neuroscience*, 19(3), 366–374.
- 1078 Rahnev, D. A., & Fleming, S. M. (2019). How experimental procedures influence estimates of  
1079 metacognitive ability. *Neuroscience of Consciousness*, 2019(1), niz009.  
1080 <http://doi.org/10.1093/nc/niz009>
- 1081 Rahnev, D. A., Koizumi, A., McCurdy, L. Y., D'Esposito, M., & Lau, H. C. (2015). Confidence Leak  
1082 in Perceptual Decision Making. *Psychological Science*, 26(11), 1664–1680.

- 1083 Rausch, M., & Zehetleitner, M. (2019). The folded X-pattern is not necessarily a statistical  
1084 signature of decision confidence. *PLoS Computational Biology*, *15*(10), e1007456.  
1085 <http://doi.org/10.1371/journal.pcbi.1007456>
- 1086 Recht, S., Mamassian, P., & de Gardelle, V. (2019). Temporal attention causes systematic biases  
1087 in visual confidence. *Scientific Reports*, *9*:11622, 1–9. [http://doi.org/10.1038/s41598-019-](http://doi.org/10.1038/s41598-019-48063-x)  
1088 [48063-x](http://doi.org/10.1038/s41598-019-48063-x)
- 1089 Sanders, J. I., Hangya, B., & Kepecs, A. (2016). Signatures of a Statistical Computation in the  
1090 Human Sense of Confidence. *Neuron*, *90*(3), 499–506.
- 1091 van den Berg, R., Zylberberg, A., Kiani, R., Shadlen, M. N., & Wolpert, D. M. (2016). Confidence is  
1092 the bridge between multi-stage decisions. *Current Biology*, *26*(23), 3157-3168.
- 1093 Wokke, M. E., Achoui, D., & Cleeremans, A. (2020). Action information contributes to  
1094 metacognitive decision-making. *Scientific Reports*, *10*: 3632, 1–15.  
1095 <http://doi.org/10.1038/s41598-020-60382-y>
- 1096 Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and  
1097 error monitoring. *Philosophical Transactions of the Royal Society of London Series B,*  
1098 *Biological Sciences*, *367*(1594), 1310–1321.



Click here to access/download  
**Supplemental Material**  
cfc\_ms\_supp.pdf

