



HAL
open science

Disentangling the origins of confidence in speeded perceptual judgments through multimodal imaging

Michael Pereira, Nathan Faivre, Iñaki Iturrate, Marco Wirthlin, Luana Serafini, Stéphanie Martin, Arnaud Desvachez, Olaf Blanke, Dimitri van de Ville, José del R. Millán

► To cite this version:

Michael Pereira, Nathan Faivre, Iñaki Iturrate, Marco Wirthlin, Luana Serafini, et al.. Disentangling the origins of confidence in speeded perceptual judgments through multimodal imaging. Proceedings of the National Academy of Sciences of the United States of America, 2020, 10.1073/pnas.1918335117 . hal-02529074

HAL Id: hal-02529074

<https://hal.science/hal-02529074>

Submitted on 3 May 2021

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.

Disentangling the origins of confidence in speeded perceptual judgments through multimodal imaging

Running title: Decision commitment improves confidence

Michael Pereira^{1,2,3*}, Nathan Faivre^{2,3,4*}, Iñaki Iturrate^{1,2*}, Marco Wirthlin^{2,3}, Luana Serafini^{2,3,5},
Stéphanie Martin^{1,2}, Arnaud Desvachez^{1,2}, Olaf Blanke^{1,2,6}, Dimitri Van De Ville^{2,7,8}, José del R.
Millán^{1,2}

Affiliations

1 Defitech Foundation Chair in Brain-Machine Interface, École Polytechnique Fédérale de Lausanne, Geneva, Switzerland

2 Center for Neuroprosthetics, École Polytechnique Fédérale de Lausanne, Geneva, Switzerland

3 Laboratory of Cognitive Neuroscience, Brain Mind Institute, Faculty of Life Sciences, École Polytechnique Fédérale de Lausanne, Geneva, Switzerland

4 Centre d'Economie de la Sorbonne, CNRS UMR 8174, Paris, France

5 Department of Biomedical, Metabolic and Neural Sciences, University of Modena and Reggio Emilia, Modena, Italy

6 Department of Neurology, University Hospital Geneva, Geneva, Switzerland

7 Medical Image Processing Lab, Institute of Bioengineering, École Polytechnique Fédérale de Lausanne, Geneva, Switzerland

8 Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland

* These authors contributed equally to this study

Corresponding author:

Michael Pereira

michael.pereira@epfl.ch

Laboratory of Cognitive Neuroscience

Campus Biotech H4

Chemin des Mines 9,

1202 Genève, Switzerland

Keywords: metacognition, error monitoring, confidence, EEG, fMRI, race modeling, inferior frontal gyrus, insula, anterior prefrontal cortex

1 Abstract

2 The human capacity to compute the likelihood that a decision is correct - known as
3 metacognition - has proven difficult to study in isolation as it usually co-occurs with decision-
4 making. Here, we isolated post-decisional from decisional contributions to metacognition by
5 combining a novel paradigm with multimodal imaging. Healthy volunteers reported their
6 confidence in the accuracy of decisions they made or decisions they observed. We found
7 better metacognitive performance for committed vs. observed decisions, indicating that
8 committing to a decision informs confidence. Relying on concurrent electroencephalography
9 and hemodynamic recordings, we found a common correlate of confidence following
10 committed and observed decisions in the inferior frontal gyrus, and a dissociation in the
11 anterior prefrontal cortex and anterior insula. We discuss these results in light of decisional
12 and post-decisional accounts of confidence, and propose a generative model of confidence
13 in which metacognitive performance naturally improves when evidence accumulation is
14 constrained upon committing a decision.

15 Introduction

16 Upon making decisions, one usually “feels” that a given choice was correct or not, which
17 allows deciding whether to commit to the choice, to seek more evidence under uncertainty,
18 or to change one’s mind and go for another option. This crucial aspect of decision making
19 relies on the capacity to monitor and report one’s own mental states, which is commonly
20 referred to as metacognitive monitoring (Fleming et al., 2012; Koriat, 2006). One promising
21 venue to unravel the neural and cognitive mechanisms of metacognitive monitoring involves
22 investigating how, and to what extent, humans become aware of their own errors (Yeung &
23 Summerfield, 2012). Typically, volunteers are asked to execute a first-order task under time
24 pressure (e.g., numerosity: which of two visual arrays contains more dots) and afterward
25 perform a second-order task by providing an estimate of confidence in their response (“how
26 sure were you that your response was correct?”). Confidence is formally defined as the
27 probability that a first-order response was correct given the available evidence (Pouget et al.,
28 2016). Distinct models have been proposed to explain how confidence is computed: some
29 models consider confidence as a fine-grained description of the same perceptual evidence
30 leading to the first-order decision (Kiani & Shadlen, 2009), sometimes enriched with post-
31 decisional processes (Pleskac et al., 2010, Van Den Berg et al., 2016; Fleming et al., 2017).
32 Other models posit that confidence stems from mechanisms different from those responsible
33 for making that decision (for review, see Grimaldi et al., 2015). However, as of today, the
34 contribution of (post-)decisional signals on confidence remains unclear, principally due to the
35 difficulty of dissociating confidence from first-order decision-making.

36 Here we combined a novel paradigm with multimodal neuroimaging to dissociate confidence
37 from decision-making. Our paradigm allowed a controlled comparison of confidence ratings
38 for decisions that were *committed* (i.e., taken and reported by participants), and decisions
39 that were merely *observed* (i.e., taken by a computer). Hereby, we could isolate the
40 contribution of decisional signals to confidence (Figure 1A). In the *active* condition, 20
41 participants were presented with two arrays of dots for 60 ms and were asked to indicate

42 which of the two arrays contained more dots by pressing a button with the left or right hand
43 (numerosity first-order task). At the end of each trial, participants had to report their
44 confidence in their response being correct or incorrect using their left hand (second-order
45 task). The *observation* condition followed the exact same procedure, except that the first-
46 order task was performed automatically: participants saw the image of a hand over the right
47 or left array of dots with identical yet shuffled timings and choice accuracy (i.e., observation
48 trials were a shuffled replay of active trials, see methods). They were then asked to report
49 their confidence in the observed decision. This allowed us to quantify metacognition for
50 committed (active condition) compared to observed (observation condition) decisions while
51 keeping perceptual evidence, first-order performance, and timing constant across conditions.
52 Both conditions were performed while recording simultaneous electroencephalography
53 (EEG) and functional magnetic resonance imaging (fMRI), to constrain blood-oxygenation
54 level dependent (BOLD) correlates of confidence to electrophysiological processes occurring
55 immediately after the committed or observed decision.

56 Data collection was conducted in view of testing three pre-registered hypotheses
57 (<https://osf.io/a5qmv>). At the behavioral level, assuming that signals associated with overt
58 decisions inform confidence judgments, we expected confidence ratings to better track first-
59 order performance for committed compared to observed decisions. Based on several
60 findings showing a role of action monitoring for confidence (e.g., Fleming & Daw, 2017;
61 Fleming et al., 2015; Faivre et al., 2018), we expected brain regions encoding confidence
62 specifically for committed decisions to be related to the cortical network involved in action
63 monitoring, and brain regions conjunctively activated in both conditions to reflect a shared
64 mechanism independent from decision commitment. Finally, we expected to find earlier
65 correlates of confidence following committed compared to observed responses, as efferent
66 information is available before visual information (Holroyd and Coles, 2002).

67

68 Results

69 **Better metacognitive performance for committed compared to observed decisions**

70 The influence of decision commitment on second-order judgments was assessed by
71 comparing metacognitive performance for committed compared to observed decisions. The
72 first-order task consisted of indicating which of two arrays contained more dots (active
73 condition), or observing a hand making that decision (observation condition) (Figure 1A). By
74 design, first-order performance was identical in the two conditions (see Methods), with an
75 average first-order accuracy of 71.2 % (\pm 1.0 %, 95 % CI, according to a 1up/2down
76 adaptive procedure), first-order response time of 385 ms \pm 8 ms, and difference of 13.1 \pm 1.7
77 dots between the two arrays.

78 We then turned to second-order performance, quantifying metacognitive performance as the
79 capacity to adapt confidence to first-order accuracy. Confidence was measured on a
80 continuous scale quantifying the probability of being correct or incorrect (ranging from 0:
81 “sure error” to 1: “sure correct”). A mixed effects logistic regression on first-order accuracy as
82 a function of confidence and condition revealed an interaction between confidence and
83 condition (model slope: odds ratios $z = 2.90$, $p = 0.004$; marginal $R^2 = 0.69$), indicating that
84 the slope between confidence and first-order accuracy was steeper in the active compared
85 to observation condition (Figure 1B). This difference in metacognitive performance was
86 present in all participants we tested, and also found when analyzing the data with tools
87 derived from second-order signal detection theory (area under the type II receiver operating
88 curve (AROC): active condition = 0.92 ± 0.02 ; observation condition = 0.90 ± 0.03 ; Wilcoxon
89 sign rank test: $V = 163$, $p = 0.03$, see SI). In addition, metacognitive performance was
90 correlated between conditions ($R^2 = 0.93$, $p < 0.001$), suggesting partially overlapping
91 mechanisms for monitoring committed and observed decisions. Of note, confidence per se
92 did not differ across conditions ($F(1,4772) = 0.01$, $p = 0.98$).

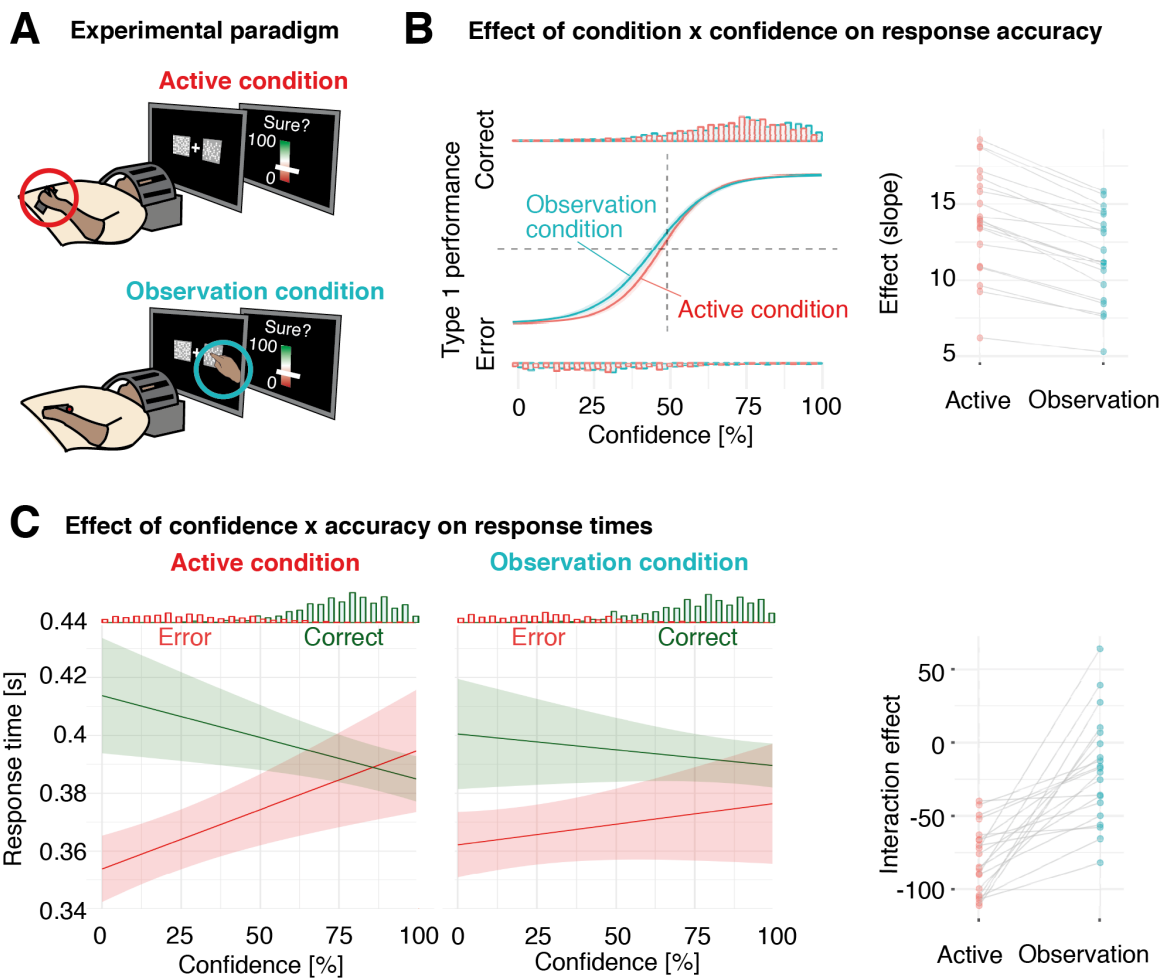
93 To assess the contribution of decisional signals to metacognitive monitoring, we ran a linear
94 mixed effects model on first-order response times as a function of confidence, accuracy, and
95 condition. This model revealed a triple interaction ($F(1,4742) = 6.05$, $p = 0.014$),
96 underscoring that in the active condition, response times for correct responses correlated
97 negatively with confidence, and response times for errors correlated positively with
98 confidence ($F(1,26) = 23.70$, $p < 0.001$, Figure 1C). No main effect of confidence ($F(1,29) =$
99 0.02 , $p = 0.89$) nor interaction between confidence and accuracy ($F(1,19) = 1.34$, $p = 0.26$)
100 was observed in the observation condition (Figure 1C). Together, these results indicate that
101 confidence was modulated by committed but not observed response times, and thus suggest
102 the importance of decisional signals and potentially motor actions to build accurate
103 confidence estimates.

104 To further elucidate the contribution of response times to confidence, we ran follow-up
105 experiments including a third condition in which the first and second-order responses were
106 reported simultaneously on a unique scale. We were able to replicate our finding of higher
107 metacognitive performance between the active and observation condition, and found that
108 metacognitive performance in the active condition was better than when first and second-
109 order responses were provided simultaneously. This confirms that the readout of speeded
110 motor actions associated with decision commitment serves subsequently as input to
111 compute confidence. Lastly, to rule out the possibility that increases in metacognitive
112 performance were due to confounding factors between the active and observed conditions
113 (e.g., demand characteristics, visual saliency), we performed the same experiment under no-
114 time pressure, and found no difference in metacognitive performance between committed
115 and observed decisions (see SI). Altogether, these results validate our first pre-registered
116 hypothesis that metacognitive performance is better for committed compared to observed
117 speeded decisions, and suggest that action monitoring might play a role in this process.

118

119

- Figure 1 -



120

121 **Figure 1. Experimental paradigm and behavioral results.** (A) Experimental paradigm: a participant
 122 lying in the fMRI bore equipped with an EEG cap performs (active condition in red) or observes
 123 (observation condition in blue) the first-order task, and subsequently reports confidence in the
 124 committed or observed decision using a visual analog scale. (B) Mixed effects logistic regression
 125 between first-order accuracy and confidence in the active (red) and observation condition (blue). The
 126 histograms represent the distributions of confidence for correct (top) and incorrect (bottom) first-order
 127 responses. Right panel: Individual slopes of the mixed effects logistic regression indicating
 128 metacognitive performance. (C) Mixed effects linear regression between first-order response times
 129 and confidence for correct (in green) and incorrect trials (in red) in the active (left panel) and
 130 observation condition (right panel). The histograms represent the distributions of response times and
 131 confidence for correct and incorrect first-order responses. Rightmost panel: interaction term between
 132 first-order accuracy and confidence for response times in the active compared to observation
 133 condition. Shaded areas represent 95% confidence intervals.

134

135

136

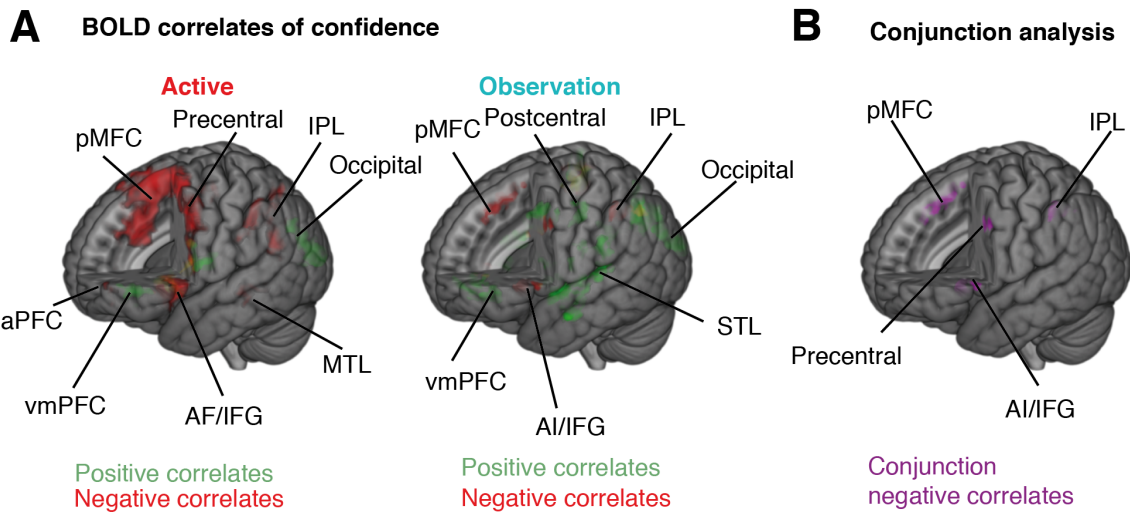
137 **BOLD correlates of confidence**

138 We sought to find the brain regions co-activating with confidence by parametrically
139 modulating a general linear model (GLM) with participants' confidence ratings, as well as
140 response times and perceptual evidence (i.e., the difference in number of dots between the
141 right and left side of the screen) as regressors of no interest (see methods). Because error
142 monitoring and confidence are tightly related (Yeung & Summerfield, 2012), we deliberately
143 analyzed the neural correlates of confidence without modeling first-order accuracy. Of note,
144 the visual scale we used allowed participants to report their confidence estimate with a
145 single and identical motor action with the left hand across conditions and trials, ruling out
146 motor confounds when analyzing data (see methods). Widespread activity correlating both
147 positively and negatively with confidence was found in the active and observation condition,
148 in line with several other studies (Fleck et al., 2005; Fleming et al., 2012b; Baird et al., 2013;
149 Heereman 2015; Hebart et al., 2016; Morales et al., 2018; Vaccaro & Fleming, 2018). A
150 complete list of activations can be found in Supplementary Table 1. In addition, we found
151 that the right precentral gyrus (contralateral to the hand reporting confidence), left insula,
152 and bilateral pMFC were significantly more predictive of confidence in the active than in the
153 observation condition (Supplementary table 2). We then defined the regions commonly
154 activated by confidence in both conditions. A conjunction analysis revealed that the bilateral
155 pMFC, left IPL, precentral gyrus, AI and IFG were negatively correlated with confidence
156 (Figure 2B; Supplementary Table 3).

157

158

- Figure 2 -



159

160 **Figure 2. BOLD correlates of confidence.** (A) Brain areas co-activated with positive (green) and negative (red)
161 confidence values for the active (left) and observation (right) conditions. (B) Brain areas co-activated with
162 negative confidence values in both conditions (conjunction analysis). All displayed BOLD activations are FWE-
163 corrected ($p < 0.05$) at the cluster-level with a threshold at $p < 0.001$. Labels: anterior insula (AI), anterior prefrontal
164 cortex (aPFC), Posterior medial frontal cortex (pMFC), inferior frontal gyrus (IFG), inferior parietal lobule (IPL),
165 medial temporal lobe (MTL), superior temporal lobe (STL), ventromedial prefrontal cortex (vmPFC). Not all brain
166 regions are labeled (see Supplementary Tables 1-3).

167

168 **ERP correlates of confidence** To further isolate the neural correlates of confidence for
169 committed and observed decisions, we identified which regions co-activated with EEG
170 correlates of confidence occurring exclusively within five hundred milliseconds after the first-
171 order response (i.e., post-decisional processes). We first modeled the EEG amplitude time-
172 locked to the first-order response as a function of confidence using mixed effects linear
173 regression, with first-order response times and perceptual evidence as covariates of no
174 interest (see methods). In the active condition, we found that EEG amplitude correlated with
175 confidence starting 68 ms following the first-order response over centro-parietal electrodes,
176 resembling a centro-parietal positivity (CPP; Figure 3A, top left; O'Connell et al., 2012).
177 Another correlate of confidence was found 88 ms post-response over frontoparietal
178 electrodes, akin to an error-related negativity (ERN; Figure 3A, bottom left; Falkenstein et al.,
179 1991, Gehring et al., 1993). In the observation condition, correlates of confidence were
180 found on the same two electrodes with similar topography (correlation between frontocentral

181 cluster in the active and observation conditions: $\rho = 0.88$) but not before 200 ms post-
182 response (Figure 3A, right).

183

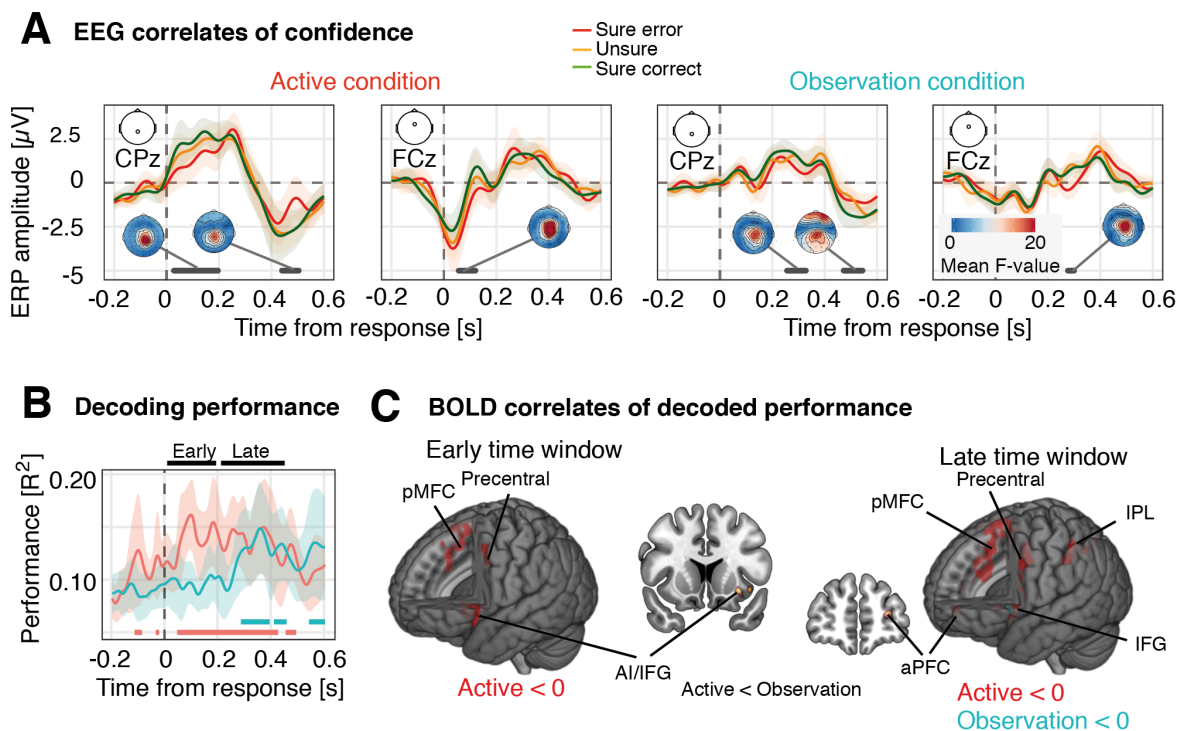
184 **Common and distinct BOLD correlates of EEG decoded confidence**

185 The brain regions corresponding to the ERP correlates of confidence were identified by
186 modeling the BOLD signal with EEG-based single-trial predictions of confidence. Confidence
187 predictions at each time point were derived from a linear regressor taking the EEG
188 independent components activation profiles as low-dimensional variables ($N=8 \pm 3$ for each
189 participant, see methods). Leave-one-out performance was significant at the group level
190 (non-parametric permutation test, corrected $p < 0.05$) with a peak decoding performance
191 achieved 96 ms and 356 ms following committed and observed responses (Figure 3B).

192 To dissociate early correlates of potentially “all-or-none”, binary error detection from fine-
193 grained second-order confidence estimates described as occurring 200 ms after response
194 (Boldt and Yeung, 2015), we selected two time points corresponding to local peaks in the
195 cross-validated decoding performance within an early (50 - 200 ms post response) and late
196 (200 - 450 ms) temporal windows (see Methods). The latency of the early peaks was $108 \pm$
197 22ms in the active condition. There was no significant decoding in the early time window in
198 the observation condition. Late peak latencies were 321 ± 31 ms in the active and 353 ± 27
199 ms in the observation condition, with no significant difference between condition ($t(19) = -$
200 1.49 , $p = 0.15$). Based on these two time-points, we re-trained one regressor per condition
201 and peak on all available epochs and used the resulting single-trial predictions as a
202 parametric regressor to model the BOLD signal, along with first-order response times and
203 perceptual evidence as covariates of no interest. By using EEG as a time-resolving proxy to
204 BOLD signal (Britz et al., 2010), we sought to investigate the anatomical correlates of
205 confidence at specific timings, with the aim of disentangling BOLD signal associated with pre
206 and post-decisional processes (Gherman & Philiastides 2018).

207

- Figure 3 -



208

209

210

211

212

213

214

215

216

217

218

219

220

221

222

223

224

225

226

Figure 3. EEG-informed correlates of confidence. (A) ERPs time-locked to the first-order response are shown for the active condition (left panels) and observation condition (right panels) for the CPz and FCz sensors. For illustrative purposes, epochs were binned according to three levels of reported confidence: sure error (0 - 33% confidence), unsure (34 - 66% confidence) and sure correct (67 - 100% confidence), although statistics were computed with raw confidence values using mixed effects linear regression. The shaded areas represent 95%-CI. Regions of significance ($p < 0.05$, FWE corrected) are depicted with a gray line, along with topographic maps of the corresponding F values. (B) Leave-one-out decoding performance over time. The plot shows the amount of variance of the reported confidence explained by the decoder (R^2) over time in the active (red trace) and the observation condition (blue trace). The shaded areas represent 95%-CI, and the horizontal dashed lines the chance level ($p < 0.05$, computed via non-parametric permutation tests corrected for multiple comparisons). For each participant and condition, the output of the best decoder within an early and late time window was retrained on the whole dataset and used as a parametric regressor to model the BOLD signal. (C) Brain areas co-activated with low decoded-confidence values in the early (left) and late time window (right). All displayed BOLD activations are FWE-corrected ($p < 0.05$) at the cluster-level with a threshold at $p < 0.001$. Labels: Posterior medial frontal cortex (pMFC), inferior parietal lobule (IPL), anterior insula (AI), inferior frontal gyrus (IFG) and anterior prefrontal cortex (aPFC). Not all brain regions are labeled (see Supplementary Table 4). The coronal view shows significant differences between the active and the observation condition for the labelled region (AI for the early time window and aPFC for the late time window).

227

228

229

230

231

232

233

The regions co-activating with *decoded* confidence in the early time window included the bilateral pMFC, the left IFG, AI and MFG (Figure 3C, left). For the late time window (Figure 3C, right), coactivations with low decoded-confidence were found in the bilateral pMFC and IFG, the left precentral gyrus, IPL, AI, MFG and aPFC for the active condition, and in the left IFG for the observation conditions (Supplementary Table 4). The left IFG was thus commonly activated by low decoded-confidence in both conditions. Differences between co-

234 activations in the active and observation condition were found in the anterior insula (AI) in
235 the early time window and in the aPFC in the late time window (Figure 3C; Supplementary
236 Table 4).

237

238 **Behavioral modeling**

239 In view of obtaining a mechanistic understanding of the way decisional and post-decisional
240 evidence contribute to confidence, we derived confidence in committed and observed
241 decisions using a race accumulator model, considered to be biologically plausible
242 representations of evidence accumulation in the brain (Bogacz et al., 2006; Gold and
243 Shadlen, 2007). Such models assume that ideal observers commit to a first-order decision
244 (D; Figure 4A) once one of two competing evidence accumulation processes (here,
245 corresponding to evidence for the left or right choice) reaches a decision.

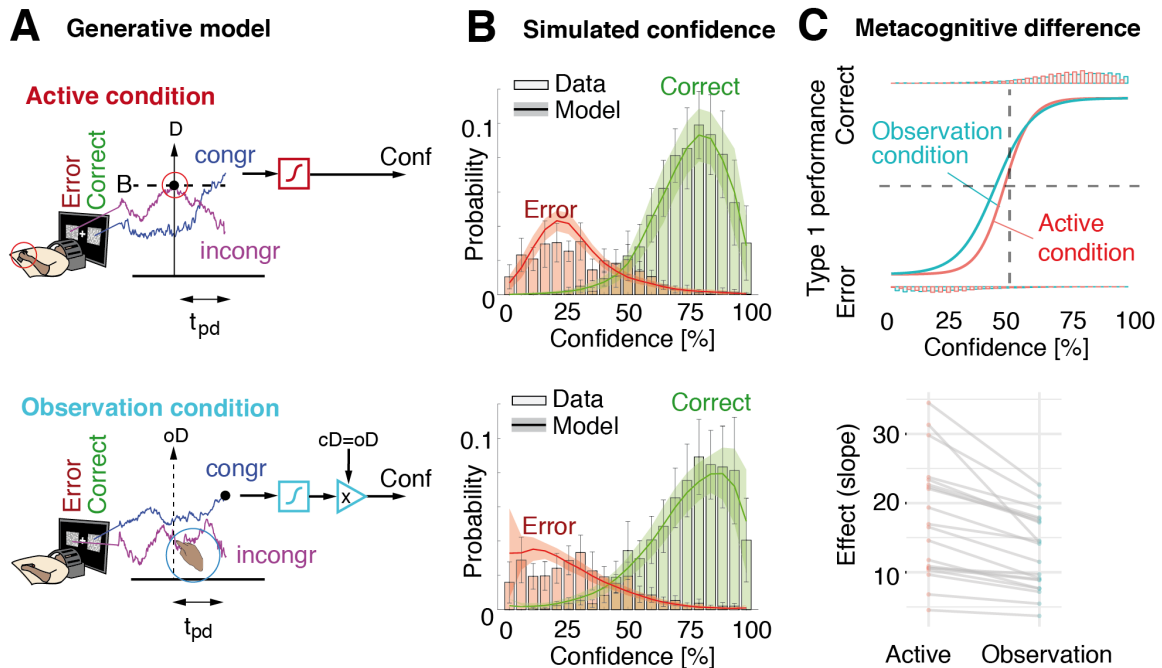
246 We first fitted five parameters (i.e., drift, bound, non-decision time, non-decision time
247 variability and starting point variability, see methods) to first-order choice accuracy and
248 response times recorded for each participant during the active condition. With these
249 parameters, we simulated pairs of competing evidence accumulation trajectories leading to
250 first-order choices and response times. We then derived confidence based on a mapping of
251 the state of evidence of the winning accumulator, following recent findings that confidence is
252 based solely on evidence supporting the decision (Peters et al. 2018; Zylberberg et al.,
253 2012). To account for changes-of-mind, we sampled accumulated evidence after a post-
254 decisional period (tpd in Figure 4A; Peskac and Busemeyer, 2010; Van Den Berg et al.,
255 2016) corresponding to the average peak decoding accuracy found with EEG (see previous
256 section). The sampled evidence was mapped to the range of confidence ratings using a
257 sigmoidal transformation with two additional free parameters controlling for bias and
258 sensitivity (see methods).

259 For the observation condition, we assumed a similar evidence accumulation process, except
260 that choice and response times were independent from the evidence accumulation process,
261 as in our paradigm. Since first-order behavior in the observation condition remained latent by
262 design, we used the parameters fitted for the active condition to simulate a second dataset
263 of pairs of competing evidence accumulation trajectories. We then mapped confidence from
264 a readout of the accumulator with highest evidence after the post-decisional period, but time-
265 locked to shuffled observed decisions (oD in Figure 4A) and response times, as in our
266 paradigm. When observed decisions were incongruent with covert decisions, we inverted the
267 simulated confidence ratings. This model fitted confidence data better than an alternative
268 model for which participants did not make covert decisions and simply readout confidence
269 from the state of evidence of the accumulator corresponding to the computer's choice (log-
270 likelihood: -2.13 ± 6.32 versus -2.91 ± 6.65 , Wilcoxon sign rank test, $p = 0.019$).

271 Across participants, our model fitted confidence ratings well (active condition: $R^2 = 0.71 \pm$
272 0.30 ; observation: $R^2 = 0.65 \pm 0.40$; Figure 4B), suggesting that it represents a plausible
273 mechanism of confidence build-up for speeded decisions. Most importantly, the confidence
274 model for the active condition predicted better metacognitive accuracy than the observation
275 model, consistent with our experimental data (Figure 4C). As in the behavioral analysis, we
276 ran a mixed effects logistic regression on first-order accuracy as a function of confidence
277 and condition, which revealed an interaction between confidence and condition (odds ratios
278 $z = -4.58$, $p < 0.001$), indicating that the slope between confidence and first-order accuracy
279 was steeper in the active compared to observation condition. Area under the type II receiver
280 operating curve (AROC) was also higher for the active condition (0.95 ± 0.02 vs. 0.93 ± 0.03 ,
281 Wilcoxon sign rank test, $V = 197$, $p < 0.001$). Of note, these differences were not explained
282 by differences in the goodness-of-fit across subjects ($R=0.13$; $p=0.59$). We could thus
283 reproduce the lower metacognitive performance found in the observation condition only by
284 detaching the decision process from the evidence accumulation process leading to
285 confidence.

286

- Figure 4 -



287

288 **Figure 4. Race accumulator model for confidence.** (A) Upper plot: an example trial for which the participant
 289 made a first-order error. The violet and blue traces represent accumulators that are incongruent and congruent
 290 with a correct response, respectively. A committed first-order decision (D) is taken when the winning accumulator
 291 hits the decision bound (dashed horizontal line). Here, the violet trace wins, producing a first-order error.
 292 Confidence is assumed to be based on the difference between both accumulators at the end of the post-
 293 decisional period. Similarly, confidence in the observed response is read-out from the difference between both
 294 accumulators at the end of the post-decisional period. In both plots, the sigmoid (square box) constrains the
 295 result to the [0,100] % interval. T_{nd} is the non-decisional time, t_d the time taken for the winning accumulator
 296 to reach the decision bound B and t_{pd} the post-decisional time. (B) Histogram of the confidence ratings obtained
 297 during the experiments, compared to the model simulations (thick line) for error (red) and correct (green). Upper
 298 plot for the active condition (second-order model), lower plot for the observation condition (non-decisional model).
 299 Error bars and shaded area represent 95% confidence intervals across subjects. (C) Top panel: Mixed logistic
 300 regression between simulated first-order accuracy and simulated confidence, in the active (red) and observation
 301 condition (blue). Bottom panel: Individual slopes of the mixed regression model indicating metacognitive
 302 performance, see Figure 1B for the actual behavioral results.

303

304 **Discussion**

305 The present study evaluated the contribution of decisional signals to metacognition by
306 comparing and modeling confidence judgments for committed and observed decisions, and
307 identifying the neural correlates of confidence with high spatiotemporal resolution. A group of
308 20 healthy volunteers was asked to perform or observe a perceptual task, and then indicate
309 their confidence regarding the accuracy of the committed or observed decisions.

310 **Better metacognitive performance for committed decisions**

311 Participants were able to adjust confidence to the accuracy of their own perceptual
312 decisions, and to the accuracy of decisions they observed. Yet, consistent with our pre-
313 registered predictions, committed decisions were associated with a slight but consistent
314 increase in metacognitive performance compared to observed decisions, which supports
315 decision commitment as an additional input for confidence. Of note, this effect could not be
316 explained by differences in terms of perceptual evidence or first-order performance across
317 conditions, which were identical by design (see Methods). A follow-up experiment revealed
318 equivalent metacognitive performance for committed and observed decisions when
319 participants were given more time to perform the first-order task. This indicates that the
320 metacognitive advantage we describe occurred in speeded tasks in which errors are
321 immediately recognized as such (Charles et al., 2013). By showing the specificity of
322 metacognitive improvement for committed decisions under speeded conditions, this follow-
323 up experiment also undermines the possibility that our effect stems from experimental
324 confounds between the active and observation conditions (e.g., demand characteristics,
325 visual saliency), as such confound would likely pertain both to speeded and non-speeded
326 conditions. Last, we found that metacognitive performance in the active condition was better
327 than another condition involving simultaneous first and second-order responses, in which by
328 definition confidence could not be informed by a previous committed decision. This brings
329 another line of evidence that action monitoring plays a role for confidence.

330 We then turned to computational modeling to shed light on the role of decisional signals for
331 decision monitoring (Kepecs et al., 2008, Kiani et al., 2009, Pleskac et al., 2010, Maniscalco
332 & Lau, 2016). One biologically plausible (computational account of decision making, called
333 race accumulator model (Bogacz et al., 2006; Kiani et al., 2014), assumes that ideal
334 observers commit to a first-order decision (here, the right or left side of the screen containing
335 more dots) once one of two competing evidence accumulation processes (for one or the
336 other choice) reaches a decision boundary. We extended these models, assuming a
337 continuation of evidence accumulation after the first-order decision (Van Den Berg et al.,
338 2016). Through this procedure, we found that the path of second-order evidence
339 accumulation in the active condition was constrained by the first-order decision boundary,
340 which translated into confidence estimates with lower variance compared to observed
341 responses which impose no constraint on evidence accumulation (7.24 ± 0.11 vs $9.04 \pm$
342 0.16 , Wilcoxon signed rank test, $V = 8$, $p < 0.001$). This prediction was verified a posteriori in
343 our behavioral data, as we found higher variance for confidence ratings in the observation
344 vs. active condition ($6.71 \% \pm 0.92$ vs. 7.33 ± 1.15 , Wilcoxon signed rank test, $V = 45$, $p =$
345 0.024).

346 The notion that committing to (but not observing) first-order decisions sharpens confidence
347 estimates is corroborated by studies showing that metacognitive performance increases
348 when response times are taken into account to compute confidence (Siedlecka, Paulewicz,
349 & Wierchoń, 2016), and decreases in case motor actions are irrelevant to the task at play
350 (Kvam et al., 2015), or when the task-relevant motor action is disrupted by transcranial
351 magnetic stimulation over premotor cortex (Fleming et al., 2015). The role of motor signals
352 for metacognition is also supported by recent results indicating that confidence increases in
353 presence of sub-threshold motor activity prior to first-order responses (Gadjos et al., 2018);
354 and that alpha desynchronization over the sensorimotor cortex controlling the hand
355 performing that action correlate with confidence (Faivre et al., 2018). Together, these
356 empirical results suggest that confidence is not solely derived from the quality of perceptual

357 evidence, but involves the perception-action cycle. By comparing committed and observed
358 decisions in a controlled way, we could test a direct prediction derived from these studies,
359 and document its neural and computational mechanisms.

360

361 **Neural correlates of confidence in committed and observed decisions**

362 After assessing the contribution of decision commitment to confidence at the behavioral
363 level, we identified the brain regions at play for monitoring committed and observed
364 decisions by parametrically modulating the BOLD signal by confidence estimates. Besides
365 brain regions activated independently across conditions (Supplementary table 1), we found
366 that the right precentral gyrus (contralateral to the hand reporting confidence), left anterior
367 insula and bilateral pMFC were significantly more predictive of confidence in the active than
368 in the observation condition (Supplementary table 2). The involvement of such motor and
369 error detection regions (Carter et al., 1998; Bonini et al., 2014; Bastin et al., 2017), together
370 with our behavioral and modeling results support the notion that action monitoring serves as
371 input for confidence. This is corroborated by behavioral results from a follow-up experiment,
372 showing that metacognitive performance was better in the active condition compared to a
373 condition in which the first and second-order responses were reported simultaneously on a
374 unique scale.

375 In search for hemodynamic correlates of confidence independent from action commitment,
376 we identified the brain regions conjunctively related to confidence in the active and
377 observation conditions as the pMFC, insula, IFG, IPL and precentral gyrus (See
378 Supplementary Table 3). This is corroborated by previous results by Heereman and
379 colleagues (2015), who found the pMFC, insula and IFG to be negatively correlated with
380 confidence during motion and color discrimination tasks, as well as Morales and colleagues
381 (2018), who found the pMFC to be negatively correlated for confidence in perceptual and
382 memory tasks. In addition, IPL activations (Hayes et al., 2011; Kim & Cabeza, 2007, 2009;

383 Moritz et al., 2006) and gray matter thickness (Filevich et al., 2018) were shown to correlate
384 negatively with confidence. These regions could represent a substrate for the computation of
385 confidence, stripped from decisional and error correction processes.

386

387 **Timing of confidence-related brain activations**

388 Due to the low temporal resolution of the BOLD signal, it is worth considering that the above-
389 mentioned regions may be contaminated by prerequisites of confidence computation (e.g.,
390 quality of numerosity representation, alertness), as well as its by-products (e.g., the act of
391 reporting confidence on the scale). To further isolate the neural correlates at play when
392 computing confidence for committed and observed decisions and pruning out some of the
393 prerequisites and by-products of confidence, we constrained our search to neural events
394 occurring in the vicinity of the committed/observed first-order response by fusing EEG and
395 fMRI data (Debener et al., 2005; Gherman & Philiastides, 2018).

396 In line with our pre-registered hypothesis, we found early correlates of confidence for
397 committed but not for observed decisions in fronto-central EEG activity resembling the error-
398 related negativity (ERN) involved in error detection (Boldt & Yeung, 2015) and in fronto-
399 parietal activity resembling the centro-parietal positivity (CPP) involved in evidence
400 accumulation (O'Connell et al., 2012). To address the possibility that early correlates of
401 confidence in observed decisions do not appear in event-related potentials but involve
402 multivariate electrophysiological patterns, we built a decoder of confidence based on whole-
403 scalp EEG. Coherently with the univariate results described above, our decoder could
404 explain confidence better than chance level in the time vicinity of committed decisions (108
405 ms post-response), while significant decoding performance was only attained 353 ms after
406 observed decisions. The absence of early correlates of confidence in the observation
407 condition was expected as participants could not possibly assess first-order accuracy before
408 perceiving the observed decision (Holroyd & Coles 2002, Van Schie et al., 2004; Iturrate et

409 al., 2015). Of note, decoding performance in the active condition plateaued after the first
410 peak and dropped after around 400 ms, indicating that ongoing processes leading to
411 confidence may be sustained in time. Thus, the computation of confidence may unfold in two
412 waves, an early one specific to the the monitoring of committed decisions, and a later one for
413 computing confidence *per se*. One possibility is that the early correlate for committed
414 decisions relates to an “all-or-none” automatic error detection system (Charles et al., 2013,
415 although see Vocat et al., 2011, Pereira et al., 2017), while the late correlate underlies a
416 fine-grained estimation of second-order signals (Boldt & Yeung, 2015).

417 We finally examined the properties of early and late correlates of confidence by assessing
418 their BOLD covariates. For that, we parametrically modulated the BOLD signal using the
419 output of a decoding model of confidence based on whole-scalp EEG, hereby obtaining a
420 time-resolved description of fMRI data (Gherman & Philiastides, 2018). In the active
421 condition, we found that the pMFC, IFG, MFG and insula were co-activated both during the
422 early and late decoding window. These regions are likely to relate to early error processing
423 based on the monitoring of errors/conflicts surrounding the first-order response (Dehaene et
424 al., 1994, Carter et al., 1998, Bonini et al., 2014, Bastin et al., 2017, Ullsperger et al., 2014
425 for a review). Furthermore, Murphy and colleagues showed that similar error-related
426 feedback signals from the pMFC inform metacognitive judgments through the modulation of
427 parietal activity involved in evidence accumulation (Murphy et al., 2015). Other regions
428 including the IPL, precentral cortex and aPFC were found specifically in the late decoding
429 window, which hints to their involvement in late processes at play for the computation of
430 graded confidence estimates. In the observation condition, the only region coactivated with
431 late electrophysiological correlates of confidence was the left IFG, adjacent to the cluster we
432 found in the active condition. This suggests the role of left IFG operating similarly around
433 300 ms whether a decision is committed or observed. Of note, the quest for domain-general
434 mechanisms of confidence (Faivre et al., 2018, Rouault et al., 2018) is hindered by the fact
435 that our paradigm alternated short blocks of active and observation conditions, which could

436 potentially inflate correlations in confidence due to confidence leaks across trials (Rahnev et
437 al., 2015).

438 By contrast to decision-independent activations in the IFG, the aPFC – commonly referred to
439 as a key region for confidence (Fleming et al., 2010, 2012, Morales et al., 2018, for review
440 see Grimaldi et al., 2015)– was involved in monitoring committed decisions only. The fact
441 that activity in the insula and aPFC were not related to confidence in observed decisions
442 reveals that these regions may underlie a putative role in linking first-order decisional signals
443 allowing early error detection to inform fine-graded confidence estimates derived from the
444 quality of perceptual evidence (Fleming et al., 2018). Beyond error detection, the aPFC
445 could operate by linking other sources of information to inform confidence, including the
446 history of confidence estimates over past trials (Shekhar et al. 2018). Although this claim
447 deserves further investigations, it extends a recent proposal by Bang & Fleming (2018)
448 arguing that aPFC is involved in reporting rather than computing confidence estimates per-
449 se.

450

451 **Conclusion**

452 We combined psychophysics, multimodal brain imaging, and computational modeling to
453 unravel the mechanisms at play when monitoring the quality of decisions we make, in
454 comparison to equivalent decisions we observe. Our behavioral and modeling results
455 indicate that committing to a decision leads to increases in metacognitive performance,
456 presumably due to the constraint of evidence accumulation by first-order decisions. By
457 focusing the analysis of neural signals on processes independent from decision-making, we
458 isolated the IFG as a key region contributing to confidence in both committed and observed
459 decisions. We further specified the functional role of the IFG, distinct from a set of regions
460 involved in error processing, and from the insula and aPFC which could potentially inform
461 confidence estimates with the output of such error processing.

462 References

- 463 Baird, B., Smallwood, J., Gorgolewski, K. J. & Margulies, D. S. Medial and lateral networks in
464 anterior prefrontal cortex support metacognitive ability for memory and perception. *J.*
465 *Neurosci.* **33**:16657–16665 (2013).
- 466 Bang, D. & Fleming, S. M. Distinct encoding of decision confidence in human medial prefrontal
467 cortex. *Proc. Natl. Acad. Sci.* **115**:6082–6087 (2018).
- 468 Bogacz, R., Brown, E., Moehlis, J., Holmes, P. & Cohen, J.D. The physics of optimal decision
469 making: A formal analysis of models of performance in two-alternative forced-choice tasks.
470 *Psychol. Rev.* **113**:700–765 (2006).
- 471 Boldt, A. & Yeung, N. Shared neural markers of decision confidence and error detection. *J.*
472 *Neurosci.* **35**:3478–3484 (2015).
- 473 Bonini, F., Burle, B., Liégeois-Chauvel, C., Régis, J., Chauvel, P., Vidal, F. Action monitoring
474 and medial frontal cortex: Leading role of supplementary motor area. *Science* **343**:888–91
475 (2014).
- 476 Britz, J., Van De Ville, D., & Michel, C. M. BOLD correlates of EEG topography reveal rapid
477 resting-state network dynamics. *Neuroimage* **52**(4), 1162-1170 (2010).
- 478 Carter, C. S., Braver, T. S., Barch, D. M., Botvinick, M. M., Noll, D. & Cohen, J. D. Anterior
479 cingulate cortex, error detection, and the online monitoring of performance. *Science*
480 **280**:747 (1998).
- 481 Charles, L., Opstal, F., Van Marti, S. & Dehaene, S. Distinct brain mechanisms for conscious
482 versus subliminal error detection. *Neuroimage* **73**:80–94 (2013).
- 483 Debener, S., et al. Trial-by-trial coupling of concurrent electroencephalogram and functional
484 magnetic resonance imaging identifies the dynamics of performance monitoring. *J.*
485 *Neurosci.* **25**(50):11730-11737 (2005).
- 486 Dehaene, S., Posner, M. I. & Tucker, D. M. Localization of a neural system for error detection
487 and compensation. *Psychol. Sci.* **5**:303–305 (1994).
- 488 Faivre, N., Filevich, E., Solovey, G., Kühn, S. & Blanke, O. Behavioural, modeling, and
489 electrophysiological evidence for supramodality in human metacognition. *J. Neurosci.*
490 **38**:0322-17 (2018).
- 491 Falkenstein, M., Hohnsbein, J., Hoormann, J., Blanke, L. Effects of crossmodal divided attention
492 on late ERP components. II. Error processing in choice reaction tasks. *Electroencephalogr.*
493 *Clin. Neurophysiol.* **78**:447–455 (1991).
- 494 Filevich, E., Forlim, C. G., Fehrman, C., Forster, C., Paulus, M., Shing, Y. L., & Kuehn, S. I know
495 that I don't know: Structural and functional connectivity underlying meta-ignorance in pre-
496 schoolers. Preprint at <https://www.biorxiv.org/content/early/2018/10/22/450346> (2018).
- 497 Fleck, M. S., Daselaar, S. M., Dobbins, I. G., & Cabeza, R. Role of prefrontal and anterior
498 cingulate regions in decision-making processes shared by memory and nonmemory tasks.
499 *Cereb. Cortex* **16**(11):1623-1630 (2005).
- 500 Fleming, S.M., Weil, R.S., Nagy, Z., Dolan, R.J. & Rees, G. Relating introspective accuracy to
501 individual differences in brain structure. *Science* **329**:1541–1543 (2010).

- 502 Fleming, S.M., Dolan, R.J. The neural basis of metacognition. *Philos. Trans. R. Soc. B. Biol. Sci.*
503 **367**:1338–1349 (2012).
- 504 Fleming, S.M., Huijgen, J & Dolan, R.J. Prefrontal contributions to metacognition in perceptual
505 decision making. *J. Neurosci.* **32**:6117–6125 (2012).
- 506 Fleming, S. M., Maniscalco, B., Ko, Y., Amendi, N., Ro, T. & Lau, H. Action-specific disruption of
507 perceptual confidence. *Psychol. Sci.* **26**:89–98 (2015).
- 508 Fleming, S. M., & Daw, N. D. Self-evaluation of decision-making: A general Bayesian framework
509 for metacognitive computation. *Psychological review*, **124**(1):91 (2017).
- 510 Fleming, S. M., Putten, E. J., & Daw, N. D. Neural mediators of changes of mind about
511 perceptual decisions. *Nature. neuroscience.* **21**:617-624 (2018).
- 512 Gajdos, T., Fleming, S., Garcia, M. S., Weindel, G. & Davranche, K. Revealing subthreshold
513 motor contributions to perceptual confidence. Preprint at
514 <https://www.biorxiv.org/content/early/2018/05/25/330605> (2018)
- 515 Gehring, W., Goss, B. & Coles, M. A neural system for error detection and compensation.
516 *Psychol. Sci.* **4**:385–390 (1993).
- 517 Gherman, S., Philiastides, M. G. Human VMPFC encodes early signatures of confidence in
518 perceptual decisions. *Elife* **7**:1–28 (2018).
- 519 Grimaldi, P., Lau, H. & Basso, M. A. There are things that we know that we know, and there are
520 things that we do not know we do not know: Confidence in decision-making. *Neurosci.*
521 *Biobehav. Rev.* **55**:88-97 (2015)
- 522 Gold, J. I. & Shadlen, M. N. The neural basis of decision making. *Annu. Rev. Neurosci.* **30**:535–
523 574 (2007).
- 524 Hayes, S. M., Buchler, N., Stokes, J., Kragel, J. & Cabeza, R. Neural correlates of confidence
525 during item recognition and source memory retrieval: evidence for both dual-process and
526 strength memory theories. *Journal of Cognitive Neuroscience* **23**(12):3959-3971 (2011).
- 527 Hebart, M. N., Schriever, Y., Donner, T. H. & Haynes, J. D. The relationship between perceptual
528 decision variables and confidence in the human brain. *Cereb. Cortex* **26**:118–130 (2016).
- 529 Heereman, J., Walter, H. & Heekeren, H. R. A task-independent neural representation of
530 subjective certainty in visual perception. *Front. Hum. Neurosci.* **9**:1–12 (2015).
- 531 Holroyd, C. B. & Coles, M. G. H. The neural basis of human error processing: Reinforcement
532 learning, dopamine, and the error-related negativity. *Psychol. Rev.* **109**:679–709 (2002).
- 533 Iturrate, I., Chavarriaga, R., Montesano, L., Minguéz, J. & Millán, J.d.R. Teaching brain-machine
534 interfaces as an alternative paradigm to neuroprosthetics control. *Sci. Rep.* **5**:13893
535 (2015).
- 536 Kepecs, A., Uchida, N., Zariwala, H. A. & Mainen, Z. F. Neural correlates, computation and
537 behavioural impact of decision confidence. *Nature* **455**:227–231 (2008).
- 538 Kiani, R. & Shadlen, M. N. Representation of confidence associated with a decision by neurons
539 in the parietal cortex. *Science* **324**:759–764 (2009).
- 540 Kiani, R., Corthell, L., & Shadlen, M. N. Choice certainty is informed by both evidence and
541 decision time. *Neuron* **84**(6):1329-1342 (2014).

- 542 Kim, H. & Cabeza, R. Trusting our memories: Dissociating the neural correlates of confidence in
543 veridical versus illusory memories. *J. Neurosci.* **27**(45):12190-12197 (2007).
- 544 Kim, H. & Cabeza, R. Common and specific brain regions in high-versus low-confidence
545 recognition memory. *Brain research* **1282**:103-113 (2009).
- 546 Koriat, A. Metacognition and consciousness In: *The Cambridge Handbook of Consciousness*,
547 289–326 (2006).
- 548 Kvam, P. D., Pleskac, T. J., Yu, S. & Busemeyer, J. R. Interference effects of choice on
549 confidence: Quantum characteristics of evidence accumulation. *Proc. Natl. Acad. Sci.*
550 **112**:10645–10650 (2015).
- 551 Maniscalco, B. & Lau, H. The signal processing architecture underlying subjective reports of
552 sensory awareness. *Neurosci. Conscious.* **1**:1–17 (2016).
- 553 Morales, J., Lau, H. & Fleming, S.M. Domain-general and domain-specific patterns of activity
554 supporting metacognition in human prefrontal cortex. *J. Neurosci.* **38**:2360–17 (2018).
- 555 Moritz, S., Gläscher, J., Sommer, T., Büchel, C. & Braus, D. F. Neural correlates of memory
556 confidence. *Neuroimage* **33**(4):1188-1193 (2006).
- 557 Murphy, P. R., Robertson, I. H., Harty, S. & O’Connell, R. G. Neural evidence accumulation
558 persists after choice to inform metacognitive judgments. *Elife* **4**:1–23 (2015).
- 559 O’Connell, R. G., Dockree, P. M. & Kelly, S.P. A supramodal accumulation-to-bound signal that
560 determines perceptual decisions in humans. *Nat. Neurosci.* **15**(12):1729-35 (2012)
- 561 Pereira, M., Sobolewski, A. & Millán, J.d.R. Action monitoring cortical activity coupled to sub-
562 movements. *eNeuro* **4**:1–12 (2017).
- 563 Peters, M. A. K. et al. Perceptual confidence neglects decision-incongruent evidence in the
564 brain. *Nat. Hum. Behav.* **1**:1–8 (2018).
- 565 Pleskac, T. J. & Busemeyer, J. R. Two-stage dynamic signal detection: A theory of choice,
566 decision time, and confidence. *Psychol. Rev.* **117**:864 (2010).
- 567 Pouget, A., Drugowitsch & J., Kepecs, A. Confidence and certainty: Distinct probabilistic
568 quantities for different goals. *Nat. Neurosci.* **19**:366–374 (2016).
- 569 Rahnev, D., Koizumi, A., McCurdy, L.Y., D’Esposito, M., Lau, H. Confidence Leak in Perceptual
570 Decision Making. *Psychol Sci* **26**:1664–1680 (2015).
- 571 Rahnev, D., Nee, D. E., Riddle, J., Larson, A.S. & D’Esposito, M. Causal evidence for frontal
572 cortex organization for perceptual decision making. *Proc. Natl. Acad. Sci.* **113**(21):6059-
573 6064 (2016).
- 574 Rouault, M., McWilliams, A., Allen, M. & Fleming, S. M. Human metacognition across domains:
575 insights from individual differences and neuroimaging. *Personality Neuroscience* **1**(e17):1-
576 13 (2018).
- 577 Shekhar, M., Rahnev, D. Distinguishing the Roles of Dorsolateral and Anterior PFC in Visual
578 Metacognition. *J. Neurosci.* **38**:5078–5087 (2018).
- 579 Siedlecka, M., Paulewicz, B. & Wierzchoń, M. But I was so sure! Metacognitive judgments are
580 less accurate given prospectively than retrospectively. *Front. Psychol.* **7**:1–8 (2016).

- 581 Ullsperger, M., Danielmeier, C. & Jocham, G. Neurophysiology of performance monitoring and
582 adaptive behavior. *Physiol. Rev.* **94**:35–79 (2014).
- 583 Vaccaro, A.G., Fleming, S.M. Thinking about thinking: A coordinate-based meta-analysis of
584 neuroimaging studies of metacognitive judgements. *Brain Neurosci Adv* **2**:1-14 (2018).
- 585 Van Den Berg, et al., A common mechanism underlies changes of mind about decisions and
586 confidence. *Elife* **5**:1–21 (2016).
- 587 Van Schie, H. T., Mars, R. B., Coles, M. G. H. & Bekkering, H. Modulation of activity in medial
588 frontal and motor cortices during error observation. *Nat. Neurosci.* **7**:549–54 (2004).
- 589 Vocat, R., Pourtois, G. & Vuilleumier, P. Parametric modulation of error-related ERP
590 components by the magnitude of visuo-motor mismatch. *Neuropsychologia* **49**:360–367
591 (2011).
- 592 Yeung, N., & Summerfield, C. Metacognition in human decision-making: Confidence and error
593 monitoring. *Phil. Trans. R. Soc. B*, **367**(1594), 1310-1321 (2012).
- 594 Zylberberg, A., Bartfeld, P., Sigman, M. The construction of confidence in a perceptual decision.
595 *Front. Integr. Neurosci.i* **6**:1–10 (2012).
- 596

597 **Methods**

598 Software and algorithms

Reagent or resource	Source	Identifier
MATLAB 2017a	Mathworks http://www.mathworks.com/products/matlab/	RRID:SCR_001622:
SPM12	https://www.fil.ion.ucl.ac.uk/spm/software/spm12/	RRID:SCR_007037
EEGLAB	http://sccn.ucsd.edu/eeglab/index.html	RRID:SCR_007292
Analyzer	BrainVision	RRID:SCR_002356
R	http://www.r-project.org/	RRID:SCR_001905
ggplot2	http://ggplot2.org/	RRID:SCR_014601
lme4	https://cran.r-project.org/web/packages/lme4/index.html	RRID:SCR_015654

599

600 CODE AVAILABILITY

601 Matlab and R code for reproducing all analyses can be found on GitHub
602 (https://gitlab.com/nfaiivre/analysis_public).

603 DATA AVAILABILITY

604 All data, analysis and modeling software scripts from this study will be made freely available
605 upon publication. Anonymized data will be stored on openneuro.org. Unthresholded
606 statistical maps can be found on NeuroVault (<https://neurovault.org/collections/4676/>)

607

608 EXPERIMENTAL MODEL AND SUBJECT DETAILS

609 The experimental paradigm, sample size, and analysis plan detailed below were registered
610 prior to data collection using the open science framework (<https://osf.io/a5qmv>).

611 Twenty-five healthy volunteers (12 females, mean age = 24.6 ± 1.43) from the student
612 population at the Swiss Federal Institute of Technology took part in this study in exchange
613 for monetary compensation (20 CHF per hour). All participants were right-handed, had
614 normal hearing and normal or corrected-to-normal vision, and no psychiatric or neurological
615 history. They were naive to the purpose of the study and gave informed consent. The study
616 was approved by the ethical committee of the canton of Geneva, Switzerland (Commission
617 Cantonale d'Ethique de la Recherche (CCER); study number 2017-00014). Five subjects
618 were excluded from the analysis: Data from three participants were not analyzed due to
619 technical issues during recording (high electrode impedance preventing data collection for
620 safety reasons), and two participants were excluded as they could not perform the first-order
621 task fast enough. The sample size was predefined based on power analyses conducted on
622 pilot data, leading to a power of 0.88 (95% CI = 0.80, 0.94) with a sample size of 25
623 participants.

624

625 METHOD DETAILS

626 **Experimental paradigm**

627 All stimuli were prepared and presented using Python 2.7. Each trial started with the display
628 of a 4° by 4° fixation cross presented for 500 to 1500 ms (uniform random distribution,
629 optimized apriori to maximize design efficiency see Friston et al., 1999). Then two square
630 boxes (size 4° by 4°) situated on each side of the fixation cross (center-to-center eccentricity
631 of 8°) were flashed for 60 ms. In total, the two boxes contained 100 dots (diameter 0.4°)
632 distributed unequally among them. Boxes and dots were displayed at maximum contrast on
633 a black background. In the active condition, participants were asked to indicate which box

634 contained most dots by pressing a key in less than 500 ms (first-order task). Responses
635 slower than 500 ms were discouraged by playing a loud alarm sound. In the observation
636 condition, participants were instructed to observe the image of a hand (6° by 6°) performing
637 the first-order task by appearing on the side of the screen corresponding to one of the two
638 boxes. They were told that the hand was controlled by a computer performing at about the
639 same level as them to discriminate the box containing most dots. Responses in the observed
640 condition corresponded to those in the active condition in a shuffled order, so that accuracy
641 and response times were kept constant across conditions (see below). After the first-order
642 response (button press or visual hand onset), a mask composed of two boxes filled with 100
643 dots each appeared in order to interrupt perceptual processing and ensure that the two
644 conditions were similar in terms of visual input. After a period of time corresponding to 2 s
645 from stimulus onset, a visual analog scale appeared instead of the mask, and participants
646 were asked to use it to report how confident they were about their own first-order response
647 (active condition), or about the observed first-order response (observation condition). The
648 scale was shown for 6.5 seconds, with marks at 0 (certainty that the first-order response was
649 erroneous), 0.5 (unsure about the first-order response) and 1.0 (certainty that the first-order
650 response was correct). A cursor moved back and forth along the scale at slow speed (3 °/s),
651 and participants had to press the left button at any moment when the cursor was at their
652 chosen confidence level. The initial position and direction of the cursor was randomized and
653 always passed through each position of the scale at least twice so that participants had one
654 more chance were they to miss the first pass of the cursor.

655 Each experimental run was divided into four blocks of 12 trials, alternating between active
656 and observation blocks. Each run started with an active block, and first-order responses in
657 that block were shuffled and replayed in the following observation block. Importantly, the
658 relation between response times, choice, and perceptual evidence was kept, as we shuffled
659 trial order only. The experiment comprised six experimental runs, totalizing 144 trials per
660 condition. During the active condition, the task difficulty was adjusted by an automatic one-

661 up two-down staircase procedure to make the first-order performance rate converge to 71%
662 (Levitt, 1971). The perceptual difficulty (defined as the difference in the number of dots
663 between the two boxes) was decreased by one after one incorrect response and increased
664 by one after two consecutive correct responses. The perceptual difficulty was pre-tuned to
665 individual perceptual abilities by performing 96 trials of the active condition without
666 confidence ratings prior to entering the scanner.

667

668 **Data collection**

669 EEG data were recorded at 5000 Hz using a 63 channel setup (BrainAmp DC-amplifier,
670 BrainProducts GmbH, Munich, Germany) synchronized to the scanner's internal clock.
671 Impedances of all channels were kept below 10K Ohms before entering the scanner. BOLD
672 signal was recorded in a 3T Prisma Siemens scanner with a 32-channel coil. We used an
673 EPI sequence (TR = 1280 ms, TE = 31 ms, FA = 64°) with 4x multiband acceleration. We
674 acquired 64 slices of 2 x 2 x 2 mm voxels without gap (FOV = 215 mm) with slice orientation
675 tilted 25° backward relative to the AC-PC line so as to include the cerebellum. Structural T1-
676 weighted images were acquired using a MPRAGE sequence (TR = 2300 ms, TE = 2.32 ms,
677 FA = 8°) with 0.9 x 0.9 x 0.9 mm voxels (FOV = 240 mm).

678

679 **QUANTIFICATION AND STATISTICAL ANALYSIS**

680 **Behavioral analysis**

681 Trials in which no first-order (2.0 %) or second-order response (2.9 %) was provided were
682 excluded. Response times (RT) were defined as the time elapsed between stimulus onset
683 and response button press (active condition), or onset of the visual hand (observation
684 condition). Trials with RT smaller than 200 ms or higher than 500 ms (due to the loud sound)
685 were also excluded from further analysis (13.1 %). Finally, trials from the observation

686 condition during which the participant mistakenly pressed the response button were also
687 excluded (12.6 %). As the exclusion criteria are not mutually exclusive, this resulted in a final
688 number of trials of 119 ± 5 trials in the active condition and 118 ± 5 trials in the observation
689 condition, out of 144 possible trials.

690 All continuous variables were analyzed using mixed effects models, using the lme4 (Bates et
691 al., 2014) and lmerTest (Kuznetsova et al., 2017) packages in R. Inclusion of random effects
692 was guided by model comparison and selection based on maximum likelihood ratio tests.
693 The significance of fixed effects was estimated using Satterthwaite's approximation for
694 degrees of freedom of F statistics (Luke 2017). All statistical tests were two-tailed.
695 Metacognitive performance was modeled using mixed effects logistic regression between
696 first-order accuracy and confidence, with random intercept for participants and random slope
697 for confidence. The slope of the model was interpreted as a metric for metacognitive
698 performance (i.e., capacity to adjust confidence based on first-order accuracy). We chose
699 this framework to analyze confidence as it is agnostic regarding the signals used to compute
700 confidence estimates (i.e., decisional compared to post-decisional locus, see Yeung &
701 Summerfield, 2015; Pleskac & Busemeyer, 2011), and the mixed model framework allows
702 analyzing raw confidence ratings even if they are unbalanced (e.g., in case participants do
703 not use all possible ratings).

704

705 **fMRI pre-processing and analysis**

706 The functional scans were realigned, resliced and normalized to MNI space using the flow
707 fields obtained by diffeomorphic anatomical registration through exponential linear algebra
708 (DARTEL; Ashburner 2007). The normalized scans were smoothed using a Gaussian kernel
709 of 5 mm full-width at half maximum (FWHM). The pre-processing was done using SPM12.
710 We modeled the BOLD signal using a general linear model (GLM) with two separate
711 regressors (stick functions at stimulus onset) for the active and observation condition as well

712 as their spatial and temporal derivatives. We then parametrically modulated the regressors
713 with three behavioral variables : the confidence ratings, the response times, and the
714 numerosity difference between the two array of dots (i.e., perceptual evidence). Bad trials as
715 defined in the behavioral analysis section were modeled by two separate regressors (one for
716 active and one for observation) and their spatial and temporal derivatives. We added six
717 realignments parameters as regressors of no interest. All second-level (group-level) results
718 are reported at a significance-level of $p < 0.05$ using cluster-extent family-wise error (FWE)
719 correction with a voxel-height threshold of $p < 0.001$. We used the anatomical automatic
720 labelling (AAL) atlas for brain parcellation (Tzourio-Mazoyer et al., 2002).

721

722 **EEG pre-processing**

723 MR-gradient artifacts were removed using sliding window average template subtraction
724 (Allen et al., 2000). TP10 electrode on the right mastoid was used to detect heartbeats for
725 ballistocardiogram artifact (BCG) removal using a semi-automatic procedure in BrainVision
726 Analyzer 2. Data were then filtered using a Butterworth, 4th order zero-phase (two-pass)
727 bandpass filter between 1 and 10 Hz, epoched [-0.2, 0.6 s] around the response onset (i.e.
728 the button press in the active condition or the appearance of the virtual hand for in
729 observation condition), re-referenced to a common average, and input to independent
730 component analysis (ICA; Makeig et al., 1996) to remove residual BCG and ocular artifacts.
731 In order to ensure numerical stability when estimating the independent components, we
732 retained 99% of the variance from the electrode space, leading to an average of 19 (SD = 6)
733 components estimated for each participant and condition. Independent components (ICs)
734 were then fitted with a dipolar source localization method (Delorme et al., 2012). ICs whose
735 dipole lied outside the brain, or resembled muscular or ocular artifacts were eliminated. A
736 total of 8 (SD = 3) components were finally kept. All preprocessing steps were performed
737 using EEGLAB and in house scripts under Matlab (The MathWorks, Inc., Natick,
738 Massachusetts, United States).

739

740 **EEG univariate analysis**

741 EEG evoked potentials were analyzed at the single trial level using a mixed effect linear
742 regression for each channel and time point. Each model included confidence or uncertainty
743 as dependent variables, with first-order response times and perceptual evidence (i.e., the
744 difference in number of dots between the right and left side of the screen) as fixed effects,
745 and a random intercept by subject. The significance of fixed effects was estimated using
746 Satterthwaite's approximation for degrees of freedom of F statistics, with family-wise error
747 correction for multiple comparisons. No random slopes were added to avoid convergence
748 failures. All analyses were performed using the tidyverse (Wickham 2017) and eegUtils
749 (Craddock, 2018) environment in R (R core team 2018).

750

751 **EEG multivariate analysis**

752 We derived a low dimensional description of the electrophysiological correlates of
753 confidence using multivariate pattern analysis on single-trials. We built independent linear
754 models in the temporal domain for each single sample within the epochs' windows, with all
755 the independent components retained as features. The models were evaluated using leave-
756 one-out cross validation to avoid overfitting, and goodness-of-fit was measured by R^2 . The
757 leave-one-out cross-validation models were also used to define the time point of maximum
758 decoding capability within two time windows of interest ([50-200] and [200-450] ms post
759 response). Once this time point was obtained for each window and participant, the
760 respective EEG values estimated from the linear regressor were fed to an EEG-fMRI
761 informed analysis (see next section).

762 Chance-level for decoding performance was computed using permutation statistics corrected
763 for multiple comparisons, by repeating the whole evaluation process 1000 times while
764 shuffling confidence rating across trials. An empirical, corrected, distribution of the null

765 hypothesis under which R^2 was not significantly different from zero was built by taking, for
766 each permutation, the maximum and minimum statistics of the R^2 throughout the whole
767 epoch window evaluated. The corrected measure of chance level was then estimated based
768 on the desired confidence of this distribution (fixed at $\alpha = 0.05$).

769

770 **EEG informed fMRI analysis**

771 To find brain-regions coactivated with decoded confidence, we built a second GLM
772 consisting of two stick function (one for each condition), parametrically modulated by four
773 variables; the output of the EEG confidence decoder at two time points post-response
774 corresponding to peak R^2 confidence decoding during the early (50 ms - 200 ms) and late
775 (200 ms - 450 ms) time windows, the response time and the numerosity difference of the
776 trial. We verified that empirical cross-correlation between regressors was low: $r_{\max} = 0.27 \pm$
777 0.05 and $r_{\max} = 0.22 \pm 0.04$ for the active and observation conditions. Excluded trials as
778 defined in the behavioral analysis section were modeled by two separate regressors (one for
779 active and one for observation) and their spatial and temporal derivatives. We added six
780 realignments parameters as regressors of no interest. All second-level (group-level) results
781 are reported at a significance-level of $p < 0.05$ using cluster-extent family-wise error (FWE)
782 correction with a voxel-height threshold of $p < 0.001$. We used the anatomical automatic
783 labelling (AAL) atlas for brain parcellation (Tzourio-Mazoyer et al., 2002).

784

785 **Behavioral modeling**

786 Our models of confidence build upon a race accumulator model predicting first-order
787 response times and choice accuracy; for every time point t (sampled at a frequency of 1000
788 Hz), each accumulator corresponded to the cumulative sum of independent draws from a
789 normal distribution with unit variance and mean equal to the drift rate (v and $-v$ for congruent
790 and incongruent choices). The decision bound was modeled as a fixed threshold B . Non-

791 decision times were modeled by a normal distribution with mean tnd and standard deviation
792 tnd_std . To model early errors, we added starting point variability; we allowed each
793 accumulator to start in a non-zero state, uniformly distributed between 0 and $zvar$ time the
794 decision bound B (Purcell & Kiani, 2016).

795 At each iteration of the optimization procedure (see below), we generated $N=1000$ surrogate
796 trials consisting in the state of the two accumulators over time and corresponding choice and
797 RT. All parameters were fitted for the active condition, through a Nelder-Mead simplex log-
798 likelihood minimization, comparing observed and simulated distribution of response times
799 with a Kolmogorov-Smirnov test. To separate correct and error trials, the sign of RT was
800 inverted for error trials. We constrained the parameters to positive values by applying an
801 exponential transformation of the variables $f(x) = \exp(x)$, except for non-decision time and
802 non-decision time variability which were constrained to $[0,1]$ s by a sigmoid transformation
803 $f(x) = 1/(1+\exp(-x))$.

804 As the state of the evidence accumulation is unconstrained, we used a second stage fitting
805 procedure to map these values to the 0-1 confidence scale. For the active condition, we
806 sampled evidence for confidence as the state of the winning accumulator at a latency
807 corresponding to peak performance in EEG decoded confidence. We divided the non-
808 decision time into a sensory and an 80ms motor component (Resulaj et al.,2008). We
809 assumed that if EEG predicted confidence best around 320 ms after the RT, then confidence
810 would depend on the state of the accumulators $320 + 80 = 400$ ms after the choice. To map
811 the evidence to a 0 - 1 confidence scale, we used a sigmoid function:

$$812 \quad C = \exp((x_1E + x_2))/(1 + \exp(x_1E + x_2)),$$

813 With C the resulting simulated confidence, E the accumulated evidence and x_1 , x_2 two
814 free parameters corresponding to the sensitivity and the bias of the mapping.

815 For the observation condition, we assumed that confidence was readout from an identical
816 evidence accumulation process, albeit disconnected from the computer's decisions (and

817 response times). We thus simulated an additional 1000 surrogate trials for the observation
818 condition but time-locked the post-decisional readout of confidence to the shuffled RTs from
819 the active condition. The confidence readout was based on the accumulator with highest
820 value, thus assuming a covert decision at the time of the read-out. We then fitted the
821 parameters of the mapping as in the active condition but inverting confidence ($c' = 1-c$)
822 when the chosen accumulator deferred from the computer's decision.

823 Additional references

- 824 Allen, P.J., Josephs, O. & Turner, R. A. Method for removing imaging artifact from continuous
825 EEG recorded during functional MRI. *Neuroimage* **239**:230–239 (2000).
- 826 Ashburner, J. A fast diffeomorphic image registration algorithm. *Neuroimage* **38**:95–113 (2007).
- 827 Bates, D., Maechler, M. Bolker, B. & Walker, S. lme4: Linear mixed-effects models using Eigen
828 and S4. R package version **1**(7), 1-23 (2014).
- 829 Craddock, M. eegUtils: A collection of utilities for EEG analysis. R package version 0.1.13.
830 <https://github.com/craddm/eegUtils> (2018)
- 831 Delorme, A., Palmer, J., Onton, J., Oostenveld, R., & Makeig, S. Independent EEG sources are
832 dipolar. *PLoS one*, **7**(2), e30135 (2012).
- 833 Friston, K. J., Zarahn, E. O. R. N. A., Josephs, O., Henson, R. N., & Dale, A. M. Stochastic
834 designs in event-related fMRI. *Neuroimage*, **10**(5):607-619 (1999).
- 835 Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. lmerTest package: tests in linear mixed
836 effects models. *Journal of Statistical Software* **82**(13) (2017).
- 837 Levitt, H. C. C. H. Transformed up-down methods in psychoacoustics. *The Journal of the*
838 *Acoustical society of America* **49**(2B): 467-477 (1971).
- 839 Luke, S. G. Evaluating significance in linear mixed-effects models in R. *Behavior Research*
840 *Methods* **49**(4), 1494-1502 (2017).
- 841 Makeig, S., Bell, A. J., Jung, T. P., & Sejnowski, T. J. Independent component analysis of
842 electroencephalographic data. *Advances in neural information processing systems* pp.
843 145-151 (1996).
- 844 Purcell, B. A. & Kiani, R. (2016). Neural mechanisms of post-error adjustments of decision policy
845 in parietal cortex. *Neuron*, **89**(3), 658-671.
- 846 R Core Team. R: A language and environment for statistical computing. R Foundation for
847 Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/> (2018).
- 848 Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. Changes of mind in decision-making.
849 *Nature*, **461**(7261), 263 (2009).
- 850 Tzourio-Mazoyer, et al., Automated anatomical labeling of activations in SPM using a
851 macroscopic anatomical parcellation of the MNI MRI single-subject
852 brain. *Neuroimage* **15**(1), 273-289 (2002).

- 853 Vickers D. Decision Processes in Visual Perception. New York, NY: Academic Press (1979).
- 854 Wickham, H. tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1.
- 855 <https://CRAN.R-project.org/package=tidyverse> (2017)
- 856

Acknowledgements: O.B. is supported by the Bertarelli Foundation, the Swiss National Science Foundation, and the European Science Foundation. D.V. is supported by the Bertarelli Foundation and the Swiss National Science Foundation. The authors are grateful to Roberto Martuzzi, Loan Mattera, Gwénaél Birot and Gisong Kim for their help during data acquisition. We thank Elisa Filevich and Roy Salomon for constructive comments on the manuscript.

Author Contributions: MP, NF, II developed the study concept and contributed to the study design. Testing and data collection were performed by MP, NF, II, AD, LS, SM, and MW. MP, NF, II and LS performed the data analysis. MP performed modeling work. MP, NF and II drafted the paper; all authors provided critical revisions and approved the final version of the paper for submission.

The authors declare no competing interests.