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

Michael Pereira[a,b,c,1,2], Nathan Faivre[a,b,c,1], Iñaki Iturrate[a,1], Marco Wirthlin[a,b], Luana Serafini[a,b,d], Stéphanie Martin[a], Arnaud Desvachez[a], Olaf Blanke[a,b,e], Dimitri Van De Ville[a,f,g], and José del R. Millán[a,h,i]

[a]Center for Neuroprosthetics, École Polytechnique Fédérale de Lausanne, 1202 Geneva, Switzerland; [b]Laboratory of Cognitive Neuroscience, Brain Mind Institute, Faculty of Life Sciences, École Polytechnique Fédérale de Lausanne, 1202 Geneva, Switzerland; [c]Laboratoire de Psychologie et Neurocognition, CNRS UMR 5105, Université Grenoble Alpes, 38400 Saint-Martin-d'Hères, France; [d]Department of Biomedical, Metabolic and Neural Sciences, University of Modena and Reggio Emilia, 41121 Modena, Italy; [e]Department of Neurology, University Hospital Geneva, 1205 Geneva, Switzerland; [f]Medical Image Processing Lab, Institute of Bioengineering, École Polytechnique Fédérale de Lausanne, 1202 Geneva, Switzerland; [g]Department of Radiology and Medical Informatics, University of Geneva, 1205 Geneva, Switzerland; [h]Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712; and [i]Department of Neurology, The University of Texas at Austin, Austin, TX 78712

Edited by Wilson S. Geisler, The University of Texas at Austin, Austin, TX, and approved March 5, 2020 (received for review October 19, 2019)

The human capacity to compute the likelihood that a decision is correct—known as metacognition—has proven difficult to study in isolation as it usually cooccurs with decision making. Here, we isolated postdecisional from decisional contributions to metacognition by analyzing neural correlates of confidence with multimodal imaging. Healthy volunteers reported their confidence in the accuracy of decisions they made or decisions they observed. We found better metacognitive performance for committed vs. observed decisions, indicating that committing to a decision may improve confidence. Relying on concurrent electroencephalography and hemodynamic recordings, we found a common correlate of confidence following committed and observed decisions in the inferior frontal gyrus and a dissociation in the anterior prefrontal cortex and anterior insula. We discuss these results in light of decisional and postdecisional accounts of confidence and propose a computational model of confidence in which metacognitive performance naturally improves when evidence accumulation is constrained upon committing a decision.

metacognition | error monitoring | confidence | EEG | fMRI

Upon making decisions, one usually "feels" that a given choice was correct or not, which allows deciding whether to commit to the choice, to seek more evidence under uncertainty, or to change one's mind and go for another option. This crucial aspect of decision making relies on the capacity to monitor and report one's own mental states, which is commonly referred to as metacognitive monitoring (1–3). One promising venue to unravel the neural and cognitive mechanisms of metacognitive monitoring involves investigating how, and to what extent, humans become aware of their own errors (4). Typically, volunteers are asked to execute a first-order task under time pressure (e.g., numerosity: which of two visual arrays contains more dots) and afterward perform a second-order task by providing an estimate of confidence in their response ("how sure were you that your response was correct?"). Confidence is formally defined as the probability that a first-order response was correct given the available evidence (5, 6). Distinct models have been proposed to explain how confidence is computed: Some models consider confidence as a fine-grained description of the same perceptual evidence leading to the first-order decision (7), sometimes enriched with postdecisional processes (8–10). Other models posit that confidence stems from mechanisms different from those responsible for making that decision (for review, see ref. 11). However, as of today, the contribution of (post)decisional signals on confidence remains unclear, principally due to the difficulty of dissociating confidence from first-order decision making.

Here we combined behavioral responses with multimodal neuroimaging to identify the driving forces of confidence judgments. Our paradigm allowed a controlled comparison of confidence ratings for decisions that were committed (i.e., taken and reported by participants) and decisions that were merely observed (i.e., taken by a computer). In the active condition, 20 participants were presented with two arrays of dots for 60 ms and were asked to indicate which of the two arrays contained more dots by pressing a button with the left or right hand (numerosity first-order task). At the end of each trial, participants had to report their confidence in their response being correct or incorrect using their left hand (second-order task). The observation condition followed the exact same procedure, except that confidence was conditional to a decision performed automatically: Participants saw the image of a hand over the right or left array of dots with identical yet shuffled timings and choice accuracy (i.e., observation trials were a shuffled replay of active trials; see *Materials and Methods*). They were then asked to report their confidence in the observed decision. This allowed us to measure confidence in committed (active condition) compared to observed (observation condition) decisions while keeping perceptual evidence and timing constant across conditions. We reasoned that confidence in the active condition

Significance

Our sense of confidence stems from the evidence leading to those decisions. This has made the study of confidence in isolation from decisional processes difficult. We devised a task in which participants rate their confidence in their own decisions or in observed decisions that are thus unrelated to decisional processes. We propose a computational account of the mechanisms underlying confidence in both conditions and reproduce participants' behavior. Furthermore, we show that activity in the inferior frontal gyrus relates to confidence, even when confidence is unrelated to decisional processes. Activity in the frontal pole and insula was more related to confidence when related to participants' own decisions. Our study provides evidence on the mechanisms and neural implementation of confidence.

Author contributions: M.P., N.F., I.I., A.D., D.V.D.V., and J.d.R.M. designed research; M.P., N.F., I.I., M.W., L.S., S.M., and A.D. performed research; M.P., N.F., I.I., M.W., and A.D. contributed new reagents/analytic tools; M.P., N.F., I.I., and L.S. analyzed data; and M.P., N.F., I.I., O.B., D.V.D.V., and J.d.R.M. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Data deposition: MATLAB and R code for reproducing all analyses can be found on GitHub (https://gitlab.com/nfaivre/eegfmri_public). Anonymized data can be found on OpenNeuro (https://openneuro.org/datasets/ds002158). Unthresholded statistical maps can be found on NeuroVault (https://neurovault.org/collections/4676/).

[1]M.P., N.F., and I.I. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: michael.pereira@univ-grenoble-alpes.fr.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1918335117/-/DCSupplemental.

First published April 1, 2020.

derives primarily from the quality of perceptual evidence and from the monitoring of action signals associated with overt decisions. In contrast, in the observation condition, confidence is still based on a (covert) decisional process but is conditioned on the observed decisions (thus requiring an additional step whereby the covert and observed decisions are compared). Therefore, confidence is orthogonalized from the decisional process and can be studied independently. These assumptions were tested with a bounded evidence accumulation model of confidence.

Both conditions were performed while recording simultaneous electroencephalography (EEG) and functional magnetic resonance imaging (fMRI), to constrain blood oxygenation level–dependent (BOLD) correlates of confidence to electrophysiological processes occurring immediately after the committed or observed decision. That is, we identified neural signals related to confidence in committed and observed decisions, independently from idiosyncratic aspects of each condition prior to the decision (e.g., motor command in the active condition, visual hand in the observation condition).

Data collection was conducted in view of testing three preregistered hypotheses (https://osf.io/a5qmv). At the behavioral level, assuming that signals associated with overt decisions inform confidence judgments, we expected confidence ratings to better track first-order performance for committed compared to observed decisions based on the same amount of perceptual evidence. Inspired by several findings showing a role of action monitoring for confidence (e.g., refs. 10 and 12–15), we expected brain regions encoding confidence specifically for committed decisions to be related to the cortical network involved in action monitoring and brain regions conjunctively activated in both conditions to reflect a shared mechanism independent from decision commitment. Finally, we expected to find earlier correlates of confidence following committed compared to observed responses, as efferent information is available before visual information (16).

Results

The influence of decision commitment on second-order judgments was assessed by comparing metacognitive performance for committed compared to observed decisions. The first-order task consisted of indicating which of two arrays contained more dots (active condition) or observing a hand making that decision (observation condition) (Fig. 1*A*). Confidence was measured on a continuous scale quantifying the probability of being correct or incorrect (ranging from 0: "sure error" to 1: "sure correct"). To make valid comparisons between confidence ratings in committed and observed decisions, we verified that first-order parameters dictating the decision were equated between conditions. By design, the amount of first-order perceptual evidence (difference of 13.1 ± 1.7 dots between the two arrays), response times (RT) (385 ms ± 8 ms), and first-order accuracy (71.2% ± 1.0%, 95% CI, according to a one-up/two-down adaptive procedure) were identical in the two conditions (*Materials and Methods*).

We then turned to second-order performance, quantifying metacognitive performance as the capacity to adapt confidence to first-order accuracy. Here we report results for the main experiment. A mixed-effects logistic regression on first-order accuracy as a function of confidence and condition revealed an interaction between confidence and condition (model slope: odds ratios z = −2.90, $P = 0.004$; marginal $R^2 = 0.69$), indicating that the slope between confidence and first-order accuracy was steeper in the active compared to the observation condition (Fig. 1*B*). Although small, this difference in metacognitive performance was present in all participants we tested (Fig. 1*B*) and also found when analyzing the data with tools derived from second-order signal detection theory (area under the type II receiver operating curve [AROC]: active condition = 0.92 ± 0.02; observation condition = 0.90 ± 0.03; Wilcoxon sign rank test: V = 163, $P = 0.03$; see *SI Appendix*). In addition, metacognitive

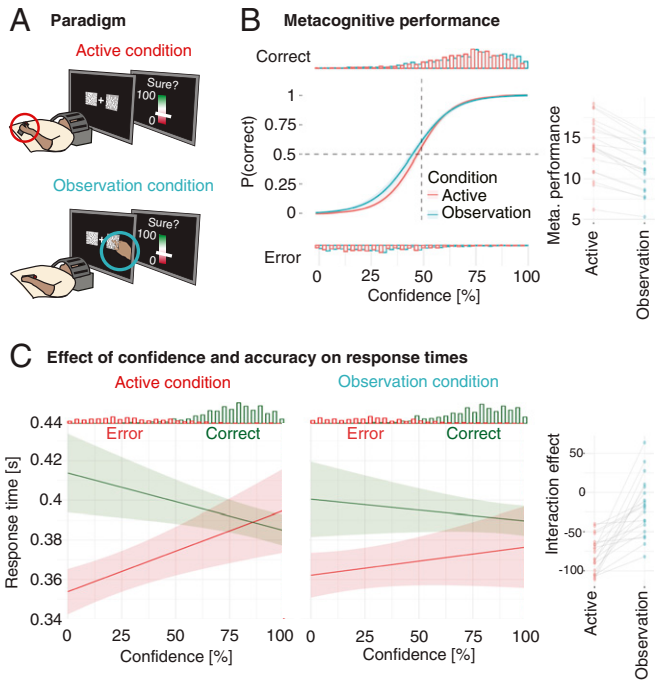

Fig. 1. Experimental paradigm and behavioral results. (*A*) Experimental paradigm: A participant lying in the fMRI bore equipped with an EEG cap performs (active condition in red) or observes (observation condition in blue) the first-order task and subsequently reports confidence in the committed or observed decision using a visual analog scale. (*B*) Mixed-effects logistic regression between first-order accuracy and confidence in the active (red) and observation condition (blue). The histograms represent the distributions of confidence for correct (*Top*) and incorrect (*Bottom*) first-order responses. (*Right*) Individual slopes of the mixed-effects logistic regression indicating metacognitive performance. (*C*) Mixed-effects linear regression between first-order RT and confidence for correct (in green) and incorrect (in red) trials in the active (*Left*) and observation condition (*Right*). The histograms represent the distributions of RT and confidence for correct and incorrect first-order responses. Rightmost: Interaction term between first-order accuracy and confidence for RT in the active compared to observation condition. Shaded areas represent 95% confidence intervals.

performance was correlated between conditions ($R^2 = 0.93$, $P < 0.001$), suggesting partially overlapping mechanisms for monitoring committed and observed decisions. Of note, confidence per se did not differ across conditions [t (19) = −0.19, $P = 0.85$, Bayes factor [BF] = 0.23].

To assess the contribution of decisional signals to metacognitive monitoring, we ran a linear mixed-effects model on first-order RT as a function of confidence, accuracy, and condition. This model revealed a triple interaction [$F(1,4742) = 6.05$, $P = 0.014$], underscoring that in the active condition, RT for correct responses correlated negatively with confidence, and response times for errors correlated positively with confidence [$F(1,26) = 23.70$, $P < 0.001$; Fig. 1*C*]. No main effect of confidence [$F(1,29) = 0.02$, $P = 0.89$] or interaction between confidence and accuracy [$F(1,19) = 1.34$, $P = 0.26$] was observed in the observation condition (Fig. 1*C*). Together, these results indicate that confidence was modulated by committed but not observed RT, and thus suggest the importance of decisional signals and potentially motor actions to build accurate confidence estimates.

To further characterize the effect of committing to a decision on metacognitive performance, we ran a first follow-up behavioral experiment comprising one session with speeded responses (under 500 ms) as in the main experiment and one session during which participants ($n = 12$) were given more time to provide their first-order response (1,500 ms; accuracy session). We also

included a third condition in both sessions, in which the first- and second-order responses were reported simultaneously on a unique scale. This allowed us to have an active condition (i.e., with a first-order response), but without requiring a commitment to a decision until the onset of the simultaneous response/confidence scale. Moreover, this condition was randomly interleaved with the observation condition so that participants did not know whether they should only observe and not respond until they saw the hand (*SI Appendix*).

In brief, this experiment revealed that the advantage of metacognitive performance for committed decisions was specific to speeded responses and did not occur when participants were given enough time to provide their first-order response (odds ratios z = 2.20, $P = 0.03$; *SI Appendix*, Fig. S1*A*). In the speeded session, we were able to replicate our finding of higher metacognitive performance between the active and observation condition (z = −2.68, $P = 0.007$) and found that metacognitive performance in the active condition was also better than when first- and second-order responses were provided simultaneously (z = −2.44, $P = 0.015$). This finding shows that the metacognitive advantage we report in the main experiment cannot be explained by a lack of attention due to the fact that participants did not have to give a first-order response.

To fully rule out the possibility that this metacognitive advantage stemmed from an attentional confound, we ran a second follow-up analysis to prove that the level of alertness/attention was equivalent between the active and observation conditions. For this, we added a visual cue in 25% of the trials in both active and observation condition (*SI Appendix*). Again, participants ($n = 14$) had higher metacognitive performance in the active condition (odds ratios z = −2.21, $P = 0.027$; *SI Appendix*, Fig. S1*B*), while there was no evidence for performance differences in detecting visual cues [average hits: 51.9% ± 14.9 in the active condition, 48.6% ± 12.9 in the observation condition, $P = 0.29$, $t(13) = 1.09$, BF = 0.45; *SI Appendix*, Fig. S1*C*]. Altogether, these results validate our first preregistered hypothesis that metacognitive performance is better for committed compared to observed speeded decisions.

Behavioral Modeling. In view of obtaining a mechanistic understanding of the way decisional and postdecisional evidence contribute to confidence, we derived confidence in committed and observed decisions using a bounded evidence accumulation model, considered to be biologically plausible representations of evidence accumulation in the brain (17, 18). Such models assume that ideal observers commit to a first-order decision (D, Fig. 2*A*) once one of two evidence accumulation processes (here, corresponding to evidence for the left or right choice) reaches a decision boundary.

We first fitted five parameters (i.e., drift, bound, nondecision time, nondecision time variability, and starting point variability; see *Materials and Methods*) to first-order choice accuracy and RT recorded for each participant during the active condition (*SI Appendix*, Fig. S2). With these parameters, we simulated pairs of evidence accumulation trajectories leading to first-order choices and RT. We then derived confidence based on a mapping of the state of evidence of the winning accumulator, following recent findings that confidence is based solely on evidence congruent with the decision (19, 20). To account for changes of mind, we sampled accumulated evidence after a postdecisional period (t_{pd} in Fig. 2*A* and refs. 8 and 9). The exact timing of this postdecisional period was taken from the EEG decoding results (*Materials and Methods*). The sampled evidence was then mapped to the range of confidence ratings using a sigmoidal transformation with two additional free parameters controlling for bias and sensitivity.

Since metacognitive performance in the active and observation conditions was highly correlated, we modeled the observation condition with a similar underlying mechanism, except that

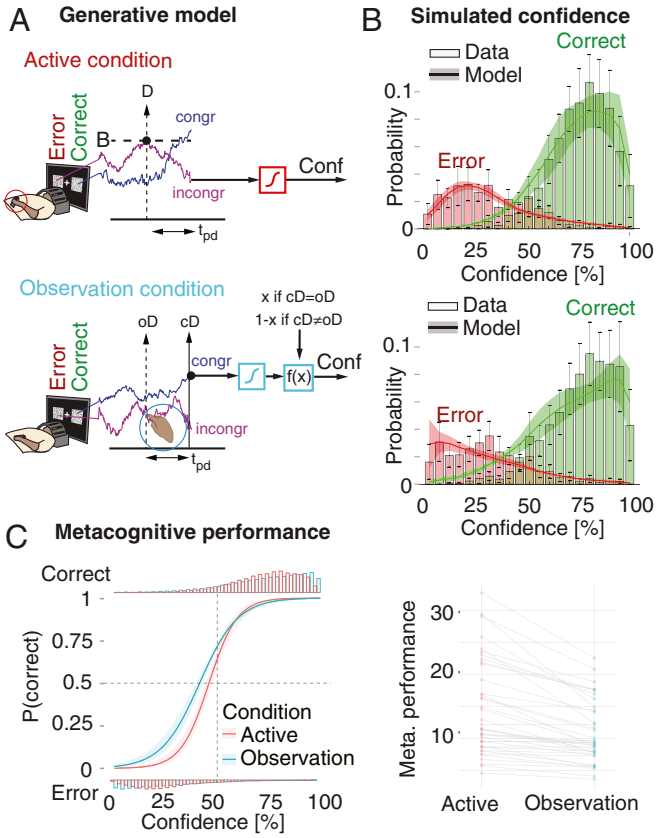

Fig. 2. Bounded evidence accumulation model for confidence. (*A*, *Upper*) An example trial for which the participant made a first-order error. The violet and blue traces represent accumulators that are incongruent and congruent with a correct response, respectively. A committed first-order decision (D) is taken when the winning accumulator hits the decision bound (dashed horizontal line). Here, the violet trace wins, producing a first-order error. Confidence is assumed to be based on the state of the accumulator corresponding to the first-order choice at the end of the postdecisional period. Confidence in the observed response is based on the state of the accumulator corresponding to the covert decision (cD) at the end of the postdecisional period, except that evidence is "inverted" if the decision cD is incongruent with the observed decision (cD ≠ oD). In both plots, the sigmoid (square box) constrains the result to the [0,100] % interval. t_{pd} is the postdecisional time. (*B*) Histogram of the confidence ratings obtained during the experiments, compared to the model simulations (thick line) for error (red) and correct (green). (*Upper*) Plot for the active condition (second-order model). (*Lower*) Plot for the observation condition (nondecisional model). Error bars and shaded area represent 95% CIs across subjects. (*C*, *Left*) Mixed logistic regression between simulated first-order accuracy and simulated confidence, in the active (red) and observation condition (blue). (*C*, *Right*) Individual slopes of the mixed regression model indicating metacognitive performance (see Fig. 1*B* for the actual behavioral results).

observed choices and RT were independent from the evidence accumulation process, as in our paradigm. Our model assumed a covert decision taken after the observed decision (oD in Fig. 2*A*). As in the active condition, confidence was defined as a readout of the winning accumulator. This (covert) confidence, however, was conditioned on the covert decision and could be related to the observed decision if the latter was congruent. Therefore, when covert and observed decisions differed, we inverted the simulated confidence. Alternative models of confidence are reported in *SI Appendix* (*SI Appendix*, Fig. S3). Across participants, our model fitted confidence ratings well (active condition: $R^2 = 0.70 \pm 0.25$; observation: $R^2 = 0.61 \pm 0.36$; Fig. 2*B* and *SI Appendix*, Figs. S4

and S5), suggesting that it represents a plausible mechanism of confidence buildup for speeded decisions. Most importantly, the confidence model for the active condition predicted better metacognitive accuracy than the observation model, consistent with our experimental data (Fig. 2*C*). As in the behavioral analysis, we ran a mixed-effects logistic regression on first-order accuracy as a function of confidence and condition, which revealed an interaction between confidence and condition (odds ratios $z = -6.01$, $P < 0.001$), indicating that the slope between confidence and first-order accuracy was steeper in the active compared to observation condition. Area under the AROC was also higher for the active condition (0.942 ± 0.006 vs. 0.921 ± 0.007, Wilcoxon sign rank test, $V = 168$, $P = 0.019$). Of note, these differences were not explained by differences in the goodness of fit across subjects ($R = 0.047$; $P = 0.85$) or by differences in readout latency between condition as using a single latency (320 ms) for all condition and subjects yielded similar results. We could thus reproduce the lower metacognitive performance found in the observation condition only by detaching the decision process from the evidence accumulation process leading to confidence.

EEG Correlates of Confidence. To isolate the neural correlates of confidence for committed and observed decisions, we identified which regions coactivated with EEG correlates of confidence occurring exclusively within 500 ms after the first-order response (i.e., postdecisional processes). We first modeled the EEG amplitude time-locked to the first-order response as a function of confidence using mixed-effects linear regression, with first-order RT and perceptual evidence as covariates of no interest (*Materials and Methods*). In the active condition, we found that EEG amplitude correlated with confidence starting 68 ms following the first-order response over centroparietal electrodes. Another correlate of confidence was found 88 ms postresponse over frontoparietal electrodes, akin to an error-related negativity (ERN; Fig. 3 *A*, bottom left and refs. 21 and 22). In the observation condition, correlates of confidence were found on the same two electrodes with similar topography (correlation between frontocentral cluster in the active and observation conditions: rho = 0.88) but not before 200-ms postresponse (Fig. 3 *A*, *Right*).

We then turned to multivariate EEG decoding to derive a single time-resolved proxy to confidence to be later used to inform the fMRI analysis. Confidence predictions at each time point were derived from a linear regressor taking the EEG-independent components activation profiles as low-dimensional variables ($n = 8 \pm 3$ for each participant; see *Materials and Methods*). Leave-one-out performance was significant at the group level (nonparametric permutation test, corrected $P < 0.05$) with a peak decoding performance achieved 96 ms and 356 ms following committed and observed responses (Fig. 3*B*; see *SI Appendix*, Fig. S7 for individual decoding performances).

To dissociate early correlates of potentially "all-or-none," binary error detection from fine-grained second-order confidence estimates described as occurring 200 ms after response (23), we selected two time points corresponding to local peaks in the cross-validated decoding performance within an early (0 to 200 ms postresponse) and late (200 to 450 ms) temporal windows (*Materials and Methods*). The latency of the early peaks was 108 ± 22 ms in the active condition. There was no significant decoding in the early time window in the observation condition. Late peak latencies were 321 ± 31 ms in the active and 353 ± 27 ms in the observation condition, with no significant difference

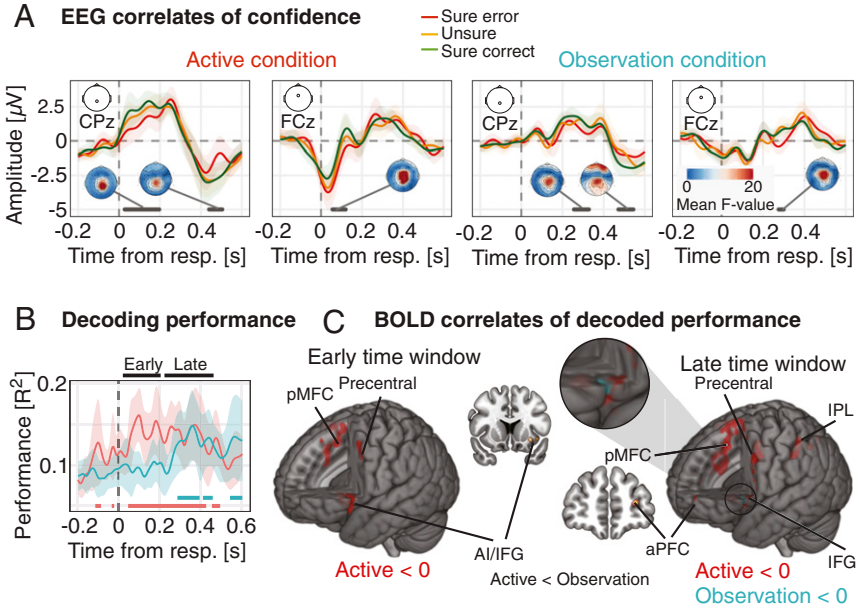

Fig. 3. EEG-informed correlates of confidence. (*A*) Event-related potentials time-locked to the first-order response (resp.) are shown for the active condition (*Left*) and observation condition (*Right*) for the CPz and FCz sensors. For illustrative purposes, epochs were binned according to three levels of reported confidence: sure error (0 to 33% confidence), unsure (34 to 66% confidence), and sure correct (67 to 100% confidence), although statistics were computed with raw confidence values using mixed-effects linear regression. The shaded areas represent 95% CI. Regions of significance ($P < 0.05$, few-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 nonparametric 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 coactivated 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 $P < 0.001$. Not all brain regions are labeled (*SI Appendix*, Table S4). The coronal view shows significant differences between the active and the observation condition for the labeled region (AI for the early time window and aPFC for the late time window).

between condition ($P = 0.20$, rank sum = 362.5). Individual latencies of the peaks in the late time window were used for the computational model. Finally, to show that our time-resolved proxy of confidence was not driven by error monitoring alone (i.e., different signal between correct and incorrect first-order responses), we confirmed that the output of the decoder correlated with confidence in correct responses only (*SI Appendix*).

Common and Distinct BOLD Correlates of EEG-Decoded Confidence. We then sought to investigate the anatomical correlates of confidence at specific timings, with the aim of disentangling BOLD signal associated with pre- and postdecisional processes. For this, we used EEG as a time-resolving proxy to the BOLD signal. Namely, we retrained one decoder for each condition and time window (i.e., at the latency corresponding to peak decoding performance), using all available epochs. We then used the resulting single-trial predictions as a parametric regressor to model the BOLD signal, along with first-order RT and perceptual evidence as covariates of no interest. The regions coactivating with decoded confidence in the early time window included the bilateral posterior medial frontal cortex (pMFC), the left inferior frontal gyrus (IFG), anterior insula (AI), and middle frontal gyrus (MFG) (Fig. 3 *C*, *Left*). For the late time window (Fig. 3 *C*, *Right*), coactivations with low decoded confidence were found in the bilateral pMFC and IFG, the left precentral gyrus, IPL, AI, MFG, and anterior prefrontal cortex (aPFC) for the active condition and in the left IFG for the observation conditions (*SI Appendix*, Table S4). The left IFG was thus commonly activated by low decoded confidence in both conditions. Differences between coactivations in the active and observation condition were found in the AI in the early time window and in the aPFC in the late time window (Fig. 3*C* and *SI Appendix*, Table S4). All of the reported activations were also found in a standard BOLD analysis with regressor parametrically modulated by confidence ratings rather than EEG-decoded confidence (*SI Appendix*, Tables S1–S3).

Discussion

The present study evaluated the origins of confidence by comparing and modeling confidence judgments for committed and observed decisions and identifying the neural correlates of confidence with high spatiotemporal resolution. A group of 20 healthy volunteers was asked to perform or observe a perceptual task and then indicate their confidence regarding the accuracy of the committed or observed decisions.

Metacognitive Performance for Committed and Observed Decisions. Participants were able to adjust confidence to the accuracy of their own perceptual decisions and to the accuracy of decisions they observed (24). Yet, consistent with our preregistered predictions, committed decisions were associated with a slight but consistent increase in metacognitive performance compared to observed decisions. Importantly, we could show that participants attended the stimulus equally well in the two conditions, as they performed similarly to detect a visual cue in a dual-task control experiment (*SI Appendix*). This indicates that the difference in metacognitive performance between conditions was not driven by differences in task demand or attention. The difference in metacognitive performance could reflect a relative decrease in the observation condition due to inherent differences in the computation of confidence. It could also reflect a relative increase in the active condition, as the monitoring of motor signals related to first-order decisions may serve to improve confidence (e.g., ref. 25). We examine these two possibilities in light of our follow-up experiments and computational model.

First, we found in a follow-up experiment that metacognitive performance for committed and observed decisions was equivalent when participants were given more time to perform the first-order task. This indicates that the metacognitive advantage (i.e., higher metacognitive performance in the active condition) we describe occurred in speeded tasks in which errors are immediately recognized as such (26). As fast error detection is based on the comparison between an action and its expected outcome (e.g., ref. 27), this result supports the existence of a relative increase of metacognitive performance in the active condition due to action monitoring. In addition, we found that metacognitive performance in the active condition was better than another condition involving simultaneous first- and second-order responses, in which by definition confidence could not be informed by a previous committed decision. This brings another line of evidence for a relative increase of metacognitive performance in the active condition, supporting the view that decision commitment plays a role for confidence. However, these data do not speak against a relative decrease of metacognitive performance in the observation condition that could coexist. The results from our follow-up experiments suggest that the cost of comparing covert and observed decisions to produce confidence estimates is relatively low, as metacognitive performance was not lower for observed vs. committed nonspeeded decisions.

We then turned to computational modeling to shed light on the potential mechanisms at play in the active and observation conditions using a bounded accumulation model (7, 17, 28, 29) assuming a continuation of evidence accumulation after the first-order decision (8, 9). Crucially, through this procedure, the path of second-order evidence accumulation in the active condition is constrained by the first-order decision boundary, which translated into confidence estimates with lower variance compared to observed responses which impose no constraint on evidence accumulation (7.07 ± 0.75 vs. 8.86 ± 1.09, Wilcoxon signed rank test, $V = 32$, $P < 0.001$). This prediction was verified a posteriori in our behavioral data, as we found higher variance for confidence ratings in the observation vs. active condition ($6.71\% \pm 0.92$ vs. 7.33 ± 1.15, Wilcoxon signed rank test, $V = 45$, $P = 0.024$). Our model reproduced first-order RT and choice accuracy in the active condition, confidence ratings in both conditions, and, importantly, the metacognitive differences observed behaviorally. This was achieved without relying on extra parameters for the observation condition. We thus favor such a parsimonious account of our data, compared to models speculating on differences in attention or difficulty between conditions that require extra parameters and are at odds with results from our follow-up experiments. Alternative models which required tuning the parameters of evidence accumulation did not improve the fit of our data.

The notion that committing to (but not observing) first-order decisions sharpens confidence estimates is corroborated by studies showing that metacognitive performance increases when RT are taken into account to compute confidence (30), and decreases in case motor actions are irrelevant to the task at play (31), or when the task-relevant motor action is disrupted by transcranial magnetic stimulation over premotor cortex (12). The role of motor signals for metacognition is also supported by recent results indicating that confidence is modulated in presence of motor activity related to first-order responses (14, 15, 32, 33). Further, alpha desynchronization over the sensorimotor cortex controlling the hand performing that action was found to correlate with confidence (13). Together, these empirical results suggest that confidence is not solely derived from the quality of perceptual evidence but involves the perception–action cycle. By comparing committed and observed decisions in a controlled way, we could test a direct prediction derived from these studies and document its neural and computational mechanisms.

Confidence-Related Brain Activations. In line with our preregistered hypothesis, we found early correlates of confidence for committed but not for observed decisions in frontocentral EEG activity resembling the ERN involved in error detection (23) and in

frontoparietal activity resembling the centroparietal positivity involved in evidence accumulation (34). To address the possibility that early correlates of confidence in observed decisions do not appear in event-related potentials but involve multivariate electrophysiological patterns, we built a decoder of confidence based on whole-scalp EEG. Coherently with the univariate results described above, our decoder could explain confidence better than chance level in the time vicinity of committed decisions (96 ms postresponse), while significant decoding performance was only attained 356 ms after observed decisions. The absence of early correlates of confidence in the observation condition was expected as participants could not possibly assess first-order accuracy before perceiving the observed decision (refs. 16, 35, and 36; see *SI Appendix*, Fig. S6 for an analysis of lateralized readiness potentials). Of note, the output of the confidence decoders still explained confidence when considering only correct responses and were thus not solely driven by dichotomic error detection (23, 26).

We then examined the neural substrate of early and late EEG correlates of confidence by assessing their BOLD covariates through the fusion of EEG and fMRI data (37). For that, we parametrically modulated the BOLD signal using the output of the confidence decoders based on whole-scalp EEG, thereby obtaining a time-resolved description of fMRI data (29). This method allowed us to constrain our search to neural events occurring in the vicinity of the committed/observed first-order response, contrary to traditional fMRI analysis in which the BOLD activity related to confidence may be contaminated by prerequisites of confidence computation (e.g., quality of numerosity representation and alertness), as well as its by-products (e.g., the act of reporting confidence on the scale). In the active condition, we found that activity in the pMFC, IFG, MFG, and insula was negatively related to decoded confidence both during the early and late decoding window. These regions are likely to relate to early error processing based on the monitoring of errors/conflicts surrounding the first-order response (refs. 38–41; see ref. 42 for a review). Furthermore, Murphy et al. (43) showed that similar error-related feedback signals from the pMFC inform metacognitive judgments through the modulation of parietal activity involved in evidence accumulation. Other regions including the IPL, precentral cortex, and aPFC were found specifically in the late decoding window, which hints at their involvement in late processes at play for the computation of graded confidence estimates (44, 45).

In the observation condition, the only region coactivated with late electrophysiological correlates of confidence was the left IFG, adjacent to the cluster we found in the active condition. This suggests the role of left IFG operating similarly around 300 ms whether a decision is committed or observed. The IFG shows common activity between action execution and observation (46). One could argue that our behavioral results are influenced by the choice of presenting a hand instead of simply highlighting the left or right stimulus. One previous study showed that disrupting the IFG impaired action understanding only when a hand was displayed (compared to a dot; ref. 47). However, in our study, the difficulty of the task does not stem from the understanding of the action but from the metacognitive evaluation of perceptual evidence. We thus favor a role of the IFG in transforming sensory evidence into confidence. Indeed, the IFG was shown to be involved for domain-general confidence (48). It is also involved in multisensory integration to form a decision (49) and its functional connectivity with sensory regions is modulated by sensory evidence (50). In our study, we show that when dissociating confidence from perceptual evidence, the IFG still tracks confidence, contrarily to other regions such as the pMFC or the aPFC.

By contrast to decision-independent activations in the IFG, activity in the aPFC—commonly referred to as a key region for confidence (51–57)—and interior insula was negatively related to confidence in committed decisions. This relation was significantly stronger in the active compared to the observed decision revealing that these regions may underlie a putative role in linking first-order decisional signals allowing early error detection to inform fine-graded confidence estimates derived from the quality of perceptual evidence (53). A recent study also found evidence that the aPFC was more activated during confidence rating than during decision making (58). Beyond error detection, the aPFC could operate by linking other sources of information to inform confidence, including the history of confidence estimates over past trials (59). Although this claim deserves further investigations, it extends a recent proposal by Bang and Fleming (60) arguing that the aPFC is involved in reporting rather than computing confidence estimates per se. All regions activated in the EEG-informed fMRI analysis were replicated in a standard fMRI analysis and are in accordance with a recent meta-analysis of confidence-related BOLD activations (61).

Conclusions

We combined psychophysics, multimodal brain imaging, and computational modeling to unravel the mechanisms at play when monitoring the quality of the decisions we make, in comparison to equivalent decisions we observe. Our behavioral and modeling results indicate that committing to a decision leads to increases in metacognitive performance, presumably due to the constraint of evidence accumulation by first-order decisions. The comparison of confidence judgments in active and observed decisions is constrained by the inherent differences that exist at the first-order level (e.g., the presence of a motor action for active but not for observation first-ordered decisions or the presence of a visual cue indicating the observed response only). We considered this comparison meaningful as all possible first-order parameters were equated between conditions. However, even under constant perceptual evidence, this comparison relies on the postulate that first-order performance itself is held constant between conditions, an assumption which remains latent as by definition it cannot be measured in the observation condition.

This is why future studies on decisional and postdecisional contributions to confidence judgments may also manipulate postdecisional evidence, leaving first-order performance intact to avoid confounding variables that may jointly influence first- and second-order behavior. Here, we mitigated the risk of having our results confounded by differences in terms of first-order processes by focusing on the neural origins of confidence after the first-order decision was made using a correlational approach. By focusing the analysis of neural signals on processes independent from decision making, we isolated two main brain regions: the IFG as a key region contributing to confidence in both committed and observed decisions and the aPFC as a region related to confidence specifically when confidence was congruent with decisional signals.

Materials and Methods

Experimental Model and Subject Details. The experimental paradigm, sample size, and analysis plan detailed below were registered prior to data collection using the Open Science Framework (https://osf.io/a5qmv).

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

Methods.

Experimental paradigm. All stimuli were prepared and presented using Python 2.7. Each trial started with the display of a 4° by 4° fixation cross presented for 500 to 1,500 ms (uniform random distribution, optimized apriori to maximize design efficiency; see ref. 62). Then two square boxes (size 4° by 4°) situated on each side of the fixation cross (center-to-center eccentricity of 8°) were flashed for 60 ms. In total, the two boxes contained 100 dots (diameter 0.4°) distributed unequally among them. Boxes and dots were displayed at maximum contrast on a black background. In the active condition, participants were asked to indicate which box contained most dots by pressing a key in less than 500 ms (first-order task). Responses slower than 500 ms were discouraged by playing a loud alarm sound. In the observation condition, participants were instructed to observe the image of a hand (6° by 6°) performing the first-order task by appearing on the side of the screen corresponding to one of the two boxes. They were told that the hand was controlled by a computer performing at about the same level as them to discriminate the box containing most dots. Responses in the observed condition corresponded to those in the active condition in a shuffled order, so that accuracy and RT were kept constant across conditions (discussed below). After the first-order response (button press or visual hand onset), a mask composed of two boxes filled with 100 dots each appeared in order to interrupt perceptual processing and ensure that the two conditions were similar in terms of visual input. After a period of time corresponding to 2 s from stimulus onset, a visual analog scale appeared instead of the mask, and participants were asked to use it to report how confident they were about their own first-order response (active condition) or about the observed first-order response (observation condition). The scale was shown for 6.5 s, with marks at 0 (certainty that the first-order response was erroneous), 0.5 (unsure about the first-order response), and 1.0 (certainty that the first-order response was correct). A cursor moved back and forth along the scale at slow speed (3°/s), and participants had to press the left button at any moment when the cursor was at their chosen confidence level. The initial position and direction of the cursor was randomized and always passed through each position of the scale at least twice so that participants had one more chance were they to miss the first pass of the cursor.

Each experimental run was divided into four blocks of 12 trials, alternating between active and observation blocks. Each run started with an active block, and first-order responses in that block were shuffled and replayed in the following observation block. Importantly, the relation between RT, choice, and perceptual evidence was kept, as we shuffled trial order only. The experiment comprised six experimental runs, totaling 144 trials per condition. During the active condition, the task difficulty was adjusted by an automatic one-up/two-down staircase procedure to make the first-order performance rate converge to 71%. The perceptual difficulty (defined as the difference in the number of dots between the two boxes) was decreased by one after one incorrect response and increased by one after two consecutive correct responses. The perceptual difficulty was pretuned to individual perceptual abilities by performing 96 trials of the active condition without confidence ratings prior to entering the scanner.

Data collection. EEG data were recorded at 5,000 Hz using a 63-channel setup (BrainAmp direct current amplifier; BrainProducts GmbH) synchronized to the scanner's internal clock. Impedances of all channels were kept below 10,000 ohms before entering the scanner. BOLD signal was recorded in a 3T Prisma Siemens scanner with a 32-channel coil. We used an echo-planar imaging sequence (repetition time [TR] = 1,280 ms, echo time [TE] = 31 ms, flip angle [FA] = 64°) with 4x multiband acceleration. We acquired 64 slices of 2- × 2- × 2-mm voxels without gap (field of view [FOV] = 215 mm) with slice orientation tilted 25° backward relative to the AC–PC line so as to include the cerebellum. Structural T1-weighted images were acquired using a MPRAGE sequence (TR = 2,300 ms, TE = 2.32 ms, FA = 8°) with 0.9- × 0.9- × 0.9-mm voxels (FOV = 240 mm).

Quantification and statistical analysis.

Behavioral analysis. Trials in which no first-order (2.0%) or second-order response (2.9%) was provided were excluded. RT were defined as the time elapsed between stimulus onset and response button press (active condition) or onset of the visual hand (observation condition). Trials with RT smaller than 200 ms or higher than 500 ms (due to the loud sound) were also excluded from further analysis (13.1%). Finally, trials from the observation condition during which the participant mistakenly pressed the response button were also excluded (12.6%). As the exclusion criteria are not mutually exclusive, this resulted in a final number of trials of 119 ± 5 trials in the active condition and 118 ± 5 trials in the observation condition, out of 144 possible trials.

All continuous variables were analyzed using mixed-effects models, using the lme4 (63) and lmerTest (64) packages in R. Inclusion of random effects was guided by model comparison and selection based on maximum likelihood ratio tests. The significance of fixed effects was estimated using Satterthwaite's approximation for degrees of freedom of F statistics (65). All statistical tests were two-tailed. Metacognitive performance was modeled using mixed-effects logistic regression between first-order accuracy and confidence, with random intercept for participants and random slope for confidence. The slope of the model was interpreted as a metric for metacognitive performance (i.e., capacity to adjust confidence based on first-order accuracy). We chose this framework to analyze confidence as it is agnostic regarding the signals used to compute confidence estimates (i.e., decisional compared to postdecisional locus; see refs. 4 and 8), and the mixed-model framework allows analyzing raw confidence ratings even if they are unbalanced (e.g., in case participants do not use all possible ratings).

Behavioral modeling. Our models of confidence build upon a bounded evidence accumulation model predicting first-order RT and choice accuracy; for every time point t (sampled at a frequency of 1,000 Hz), each accumulator corresponded to the cumulative sum of independent draws from a normal distribution with unit variance and mean equal to the drift rate (v and −v for congruent and incongruent choices). The decision bound was modeled as a fixed threshold B. Nondecision times were modeled by a normal distribution with mean tnd and SD tnd_{std}. To model early errors, we added starting point variability; we allowed each accumulator to start in a nonzero state, uniformly distributed between 0 and zvar times the decision bound B.

At each iteration of the optimization procedure (discussed below), we generated $n = 1{,}000$ surrogate trials consisting in the state of the two accumulators over time and corresponding choice and RT. All parameters were fitted for the active condition, through a Nelder–Mead simplex log-likelihood maximization, comparing observed and simulated distribution of RT with a Kolmogorov–Smirnov test. To separate correct and error trials, the sign of RT was inverted for error trials. We constrained the parameters to positive values by applying an exponential transformation of the variables $f(x) = \exp(x)$, except for nondecision time and nondecision time variability which were constrained to [0,1] s by a sigmoid transformation $f(x) = 1/(1 + \exp(-x))$. We repeated the procedure with values of zvar between 0 and 1 (steps of 0.1) and choose the model resulting in the lowest log likelihood.

As the state of the evidence accumulation is unconstrained, we used a second stage fitting procedure to map these values to the 0-to-1 confidence scale. For the active condition, we sampled evidence for confidence as the state of the winning accumulator ($n = 1{,}000$) at a latency corresponding to individual peak performance in EEG-decoded confidence plus an 80-ms motor component corresponding to the time between the decision and the actual motor response (66). To map the evidence to a 0-to-1 confidence scale, we used a sigmoid function:

$$C = \exp((X_1 E + X_2))/(1 + \exp(X_1 E + X_2)),$$

with C the resulting simulated confidence, E the accumulated evidence, and X_1, X_2 two free parameters corresponding to the sensitivity and the bias of the mapping.

For the observation condition, we assumed that confidence was readout from an identical evidence accumulation process, albeit disconnected from the computer's decisions (and RT). We thus simulated an additional 1,000 surrogate trials for the observation condition but time-locked the postdecisional readout of confidence to the shuffled RTs from the active condition. The confidence readout was based on the accumulator with highest value, thus assuming a covert decision at the time of the readout. We then fitted the parameters of the mapping as in the active condition but inversing confidence ($c' = 1 - c$) when the chosen accumulator deferred from the computer's decision.

EEG preprocessing. MR-gradient artifacts were removed using sliding window average template subtraction (67). The TP10 electrode on the right mastoid was used to detect heartbeats for ballistocardiogram artifact removal using a semiautomatic procedure in BrainVision Analyzer 2. Data were then filtered using a Butterworth, fourth-order zero-phase (two-pass) bandpass filter between 1 and 10 Hz, epoched [−0.2, 0.6 s] around the response onset (i.e., the button press in the active condition or the appearance of the virtual hand in the observation condition), rereferenced to a common average, and input to independent component (IC) analysis (68) to remove residual ballistocardiogram and ocular artifacts. In order to ensure numerical stability when estimating the independent components, we retained 99% of the variance from the electrode space, leading to an average of 19 (SD = 6) components estimated for each participant and condition. ICs were then fitted with a dipolar source localization method (69). ICs whose dipole lied outside the brain, or resembled muscular or ocular artifacts were eliminated. A total of 8 (SD = 3) components were finally kept. All preprocessing steps were performed using EEGLAB and in-house scripts under MATLAB (The MathWorks, Inc.).

EEG univariate analysis. EEG-evoked potentials were analyzed at the single-trial level using a mixed-effect linear regression for each channel and time point. Each model included confidence or uncertainty as dependent variables, with first-order RT and perceptual evidence (i.e., the difference in number of dots between the right and left side of the screen) as fixed effects, and a random intercept by subject. The significance of fixed effects was estimated using Satterthwaite's approximation for degrees of freedom of F statistics, with familywise error correction for multiple comparisons. No random slopes were added to avoid convergence failures. All analyses were performed using the tidyverse (70) environment in R (R Core Team).

EEG multivariate analysis. We derived a low-dimensional description of the electrophysiological correlates of confidence using multivariate pattern analysis on single trials. We built independent linear models in the temporal domain for each single sample within the epochs' windows, with all of the ICs retained as features. The models were evaluated using leave-one-out cross-validation to avoid overfitting, and goodness of fit was measured by R^2. The leave-one-out cross-validation models were also used to define the time point of maximum decoding capability within two time windows of interest ([0 to 200] and [200 to 450] ms postresponse). Once this time point was obtained for each window and participant, the respective EEG values estimated from the linear regressor were fed to an EEG–fMRI informed analysis (discussed in the next section).

Chance level for decoding performance was computed using permutation statistics corrected for multiple comparisons, by repeating the whole evaluation process 1,000 times while shuffling confidence rating across trials. An empirical, corrected distribution of the null hypothesis under which R^2 was not significantly different from zero was built by taking, for each permutation, the maximum statistics of the R^2 throughout the whole epoch window evaluated. The corrected measure of chance level was then estimated based on the desired confidence of this distribution (fixed at = 0.05).

EEG-informed fMRI analysis. The functional scans were realigned, resliced, and normalized to Montreal Neurological Institute space using the flow fields obtained by diffeomorphic anatomical registration through exponential linear algebra (DARTEL; ref. 71). The normalized scans were smoothed using a Gaussian kernel of 5 mm full width at half maximum. The preprocessing was done using SPM12. To find brain regions coactivated with decoded confidence, we built a generalized linear model consisting of two stick functions (one for each condition), parametrically modulated by four variables: the output of the EEG confidence decoder at two time points postresponse corresponding to peak R^2 confidence decoding during the early (0 to 200 ms) and late (200 to 450 ms) time windows, the RT, and the numerosity difference of the trial. Empirical cross-correlation between regressors confirmed limited colinearity for both the active (max(abs(R)) = 0.31 ± 0.02) and observation condition (max(abs(R)) = 0.27 ± 0.02). Excluded trials as defined in the behavioral analysis section were modeled by two separate regressors (one for active and one for observation) and their spatial and temporal derivatives. We added six realignments parameters as regressors of no interest. All second-level (group-level) results are reported at a significance-level of $P < 0.05$ using cluster-extent familywise error (FWE) correction with a voxel-height threshold of $P < 0.001$. We used the anatomical automatic labeling atlas for brain parcellation (72).

ACKNOWLEDGMENTS. O.B. is supported by the Bertarelli Foundation, the Swiss National Science Foundation, and the European Science Foundation. D.V.D.V. is supported by the Bertarelli Foundation and the Swiss National Science Foundation. N.F. has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme grant 803122. We thank Roberto Martuzzi, Loan Mattera, Gwénaël Birot, Gisong Kim, and Léa Vidal for their help during data acquisition and Elisa Filevich, Roy Salomon, and two anonymous reviewers for constructive comments on the manuscript.

1. S. M. Fleming, R. J. Dolan, The neural basis of metacognitive ability. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **367**, 1338–1349 (2012).
2. F. Meyniel, M. Sigman, Z. F. Mainen, Confidence as Bayesian probability: From neural origins to behavior. *Neuron* **88**, 78–92 (2015).
3. A. Koriat, "Metacognition and consciousness" in *The Cambridge Handbook of Consciousness*, P. D. Zelazo, M. Moscovitch, E. Thompson, Eds. (Cambridge University Press, 2007), 289–326.
4. N. Yeung, C. Summerfield, Metacognition in human decision-making: Confidence and error monitoring. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **367**, 1310–1321 (2012).
5. A. Pouget, J. Drugowitsch, A. Kepecs, Confidence and certainty: Distinct probabilistic quantities for different goals. *Nat. Neurosci.* **19**, 366–374 (2016).
6. J. I. Sanders, B. Hangya, A. Kepecs, Signatures of a statistical computation in the human sense of confidence. *Neuron* **90**, 499–506 (2016).
7. R. Kiani, M. N. Shadlen, Representation of confidence associated with a decision by neurons in the parietal cortex. *Science* **324**, 759–764 (2009).
8. T. J. Pleskac, J. R. Busemeyer, Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychol. Rev.* **117**, 864–901 (2010).
9. R. van den Berg *et al*., A common mechanism underlies changes of mind about decisions and confidence. *eLife* **5**, e12192 (2016).
10. S. M. Fleming, N. D. Daw, Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychol. Rev.* **124**, 91–114 (2017).
11. P. Grimaldi, H. Lau, M. A. Basso, There are things that we know that we know, and there are things that we do not know we do not know: Confidence in decision-making. *Neurosci. Biobehav. Rev.* **55**, 88–97 (2015).
12. S. M. Fleming *et al*., Action-specific disruption of perceptual confidence. *Psychol. Sci.* **26**, 89–98 (2015).
13. N. Faivre, E. Filevich, G. Solovey, S. Kühn, O. Blanke, Behavioural, modeling, and electrophysiological evidence for supramodality in human metacognition. *J. Neurosci.* **38**, 263–277 (2018).
14. N. Faivre *et al*., Confidence in perceptual decision-making is preserved in schizophrenia. medRxiv:10.1101/2019.12.15.19014969 (18 December 2019).
15. E. Filevich, C. Koß, N. Faivre, Response-related signals increase confidence but not metacognitive performance. bioRxiv:10.1101/735712 (15 August 2019).
16. C. B. Holroyd, M. G. H. Coles, The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychol. Rev.* **109**, 679–709 (2002).
17. R. Bogacz, E. Brown, J. Moehlis, P. Holmes, J. D. Cohen, The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol. Rev.* **113**, 700–765 (2006).
18. J. I. Gold, M. N. Shadlen, The neural basis of decision making. *Annu. Rev. Neurosci.* **30**, 535–574 (2007).
19. M. A. K. Peters *et al*., Perceptual confidence neglects decision-incongruent evidence in the brain. *Nat. Hum. Behav.* **1**, 1–8 (2017).
20. A. Zylberberg, P. Barttfeld, M. Sigman, The construction of confidence in a perceptual decision. *Front. Integr. Neurosci.* **6**, 79 (2012).
21. M. Falkenstein, J. Hohnsbein, J. Hoormann, L. Blanke, Effects of crossmodal divided attention on late ERP components. II. Error processing in choice reaction tasks. *Electroencephalogr. Clin. Neurophysiol.* **78**, 447–455 (1991).
22. W. Gehring, B. Goss, M. Coles, A neural system for error detection and compensation. *Psychol. Sci.* **4**, 385–390 (1993).
23. A. Boldt, N. Yeung, Shared neural markers of decision confidence and error detection. *J. Neurosci.* **35**, 3478–3484 (2015).
24. S. R. Patel *et al*., Single-neuron responses in the human nucleus accumbens during a financial decision-making task. *J. Neurosci.* **32**, 7311–7315 (2012).
25. R. Kiani, L. Corthell, M. N. Shadlen, Choice certainty is informed by both evidence and decision time. *Neuron* **84**, 1329–1342 (2014).
26. L. Charles, F. Van Opstal, S. Marti, S. Dehaene, Distinct brain mechanisms for conscious versus subliminal error detection. *Neuroimage* **73**, 80–94 (2013).
27. M. G. H. Coles, M. K. Scheffers, C. B. Holroyd, Why is there an ERN/Ne on correct trials? Response representations, stimulus-related components, and the theory of error-processing. *Biol. Psychol.* **56**, 173–189 (2001).
28. A. Kepecs, N. Uchida, H. A. Zariwala, Z. F. Mainen, Neural correlates, computation and behavioural impact of decision confidence. *Nature* **455**, 227–231 (2008).
29. S. Gherman, M. G. Philiastides, Human VMPFC encodes early signatures of confidence in perceptual decisions. *eLife* **7**, e38293 (2018).
30. M. Siedlecka, B. Paulewicz, M. Wierzchoń, But I was so sure! Metacognitive judgments are less accurate given prospectively than retrospectively. *Front. Psychol.* **7**, 218 (2016).
31. P. D. Kvam, T. J. Pleskac, S. Yu, J. R. Busemeyer, Interference effects of choice on confidence: Quantum characteristics of evidence accumulation. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 10645–10650 (2015).
32. T. Gajdos, S. M. Fleming, M. Saez Garcia, G. Weindel, K. Davranche, Revealing subthreshold motor contributions to perceptual confidence. *Neurosci. Conscious.* **2019**, niz001 (2019).
33. D. Dotan, F. Meyniel, S. Dehaene, On-line confidence monitoring during decision making. *Cognition* **171**, 112–121 (2018).
34. R. G. O'Connell, P. M. Dockree, S. P. Kelly, A supramodal accumulation-to-bound signal that determines perceptual decisions in humans. *Nat. Neurosci.* **15**, 1729–1735 (2012).
35. H. T. van Schie, R. B. Mars, M. G. H. Coles, H. Bekkering, Modulation of activity in medial frontal and motor cortices during error observation. *Nat. Neurosci.* **7**, 549–554 (2004).
36. I. Iturrate, R. Chavarriaga, L. Montesano, J. Minguez, J. d. R. Millán, Teaching brain-machine interfaces as an alternative paradigm to neuroprosthetics control. *Sci. Rep.* **5**, 13893 (2015).
37. S. Debener *et al*., Trial-by-trial coupling of concurrent electroencephalogram and functional magnetic resonance imaging identifies the dynamics of performance monitoring. *J. Neurosci.* **25**, 11730–11737 (2005).
38. S. Dehaene, M. I. Posner, D. M. Tucker, Localization of a neural system for error detection and compensation. *Psychol. Sci.* **5**, 303–305 (1994).
39. C. S. Carter *et al*., Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science* **280**, 747–749 (1998).
40. F. Bonini *et al*., Action monitoring and medial frontal cortex: Leading role of supplementary motor area. *Science* **343**, 888–891 (2014).
41. J. Bastin *et al*., Direct recordings from human anterior insula reveal its leading role within the error-monitoring network. *Cereb. Cortex* **27**, 1545–1557 (2017).
42. M. Ullsperger, C. Danielmeier, G. Jocham, Neurophysiology of performance monitoring and adaptive behavior. *Physiol. Rev.* **94**, 35–79 (2014).
43. P. R. Murphy, I. H. Robertson, S. Harty, R. G. O'Connell, Neural evidence accumulation persists after choice to inform metacognitive judgments. *eLife* **4**, e11946 (2015).

44. M. N. Hebart, Y. Schriever, T. H. Donner, J. D. Haynes, The relationship between perceptual decision variables and confidence in the human brain. *Cereb. Cortex* **26**, 118–130 (2016).
45. J. Muraskin *et al*., A multimodal encoding model applied to imaging decision-related neural cascades in the human brain. *Neuroimage* **180**, 211–222 (2018).
46. G. Pobric, A. F. Hamilton, Action understanding requires the left inferior frontal cortex. *Curr. Biol.* **16**, 524–529 (2006).
47. N. Nishitani, R. Hari, Temporal dynamics of cortical representation for action. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 913–918 (2000).
48. J. Morales, H. Lau, S. M. Fleming, Domain-general and domain-specific patterns of activity supporting metacognition in human prefrontal cortex. *J. Neurosci.* **38**, 3534–3546 (2018).
49. U. Noppeney, D. Ostwald, S. Werner, Perceptual decisions formed by accumulation of audiovisual evidence in prefrontal cortex. *J. Neurosci.* **30**, 7434–7446 (2010).
50. F. Filimon, M. G. Philiastides, J. D. Nelson, N. A. Kloosterman, H. R. Heekeren, How embodied is perceptual decision making? Evidence for separate processing of perceptual and motor decisions. *J. Neurosci.* **33**, 2121–2136 (2013).
51. S. M. Fleming, R. S. Weil, Z. Nagy, R. J. Dolan, G. Rees, Relating introspective accuracy to individual differences in brain structure. *Science* **329**, 1541–1543 (2010).
52. S. M. Fleming, J. Huijgen, R. J. Dolan, Prefrontal contributions to metacognition in perceptual decision making. *J. Neurosci.* **32**, 6117–6125 (2012).
53. S. M. Fleming, E. J. van der Putten, N. D. Daw, Neural mediators of changes of mind about perceptual decisions. *Nat. Neurosci.* **21**, 617–624 (2018).
54. B. Baird, J. Smallwood, K. J. Gorgolewski, D. S. Margulies, Medial and lateral networks in anterior prefrontal cortex support metacognitive ability for memory and perception. *J. Neurosci.* **33**, 16657–16665 (2013).
55. L. Qiu *et al*., The neural system of metacognition accompanying decision-making in the prefrontal cortex. *PLoS Biol.* **16**, e2004037 (2018).
56. D. Rahnev, D. E. Nee, J. Riddle, A. S. Larson, M. D'Esposito, Causal evidence for frontal cortex organization for perceptual decision making. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 6059–6064 (2016).
57. M. Rouault, A. McWilliams, M. G. Allen, S. M. Fleming, Human metacognition across domains: Insights from individual differences and neuroimaging. *Personal. Neurosci.* **1**, e17 (2018).
58. J. Yeon, D. Rahnev, Overlapping and unique neural circuits support perceptual decision making and confidence. bioRxiv:10.1101/439463 (10 October 2018).
59. M. Shekhar, D. Rahnev, Distinguishing the roles of dorsolateral and anterior PFC in visual metacognition. *J. Neurosci.* **38**, 5078–5087 (2018).
60. D. Bang, S. M. Fleming, Distinct encoding of decision confidence in human medial prefrontal cortex. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 6082–6087 (2018).
61. A. G. Vaccaro, S. M. Fleming, Thinking about thinking: A coordinate-based meta-analysis of neuroimaging studies of metacognitive judgements. *Brain Neurosci. Adv.* **2**, 2398212818810591 (2018).
62. K. J. Friston *et al*., Stochastic designs in event-related fMRI. *Neuroimage* **10**, 607–619 (1999).
63. D. Bates, M. Mächler, B. Bolker, S. Walker, Fitting linear mixed-effects models using lme4. *J. Stat. Softw.*, 10.18637/jss.v067.i01 (2015).
64. A. Kuznetsova, P. B. Brockhoff, R. H. B. Christensen, lmerTest package: Tests in linear mixed effects models. *J. Stat. Softw.*, 10.18637/jss.v082.i13 (2017).
65. S. G. Luke, Evaluating significance in linear mixed-effects models in R. *Behav. Res. Methods* **49**, 1494–1502 (2017).
66. A. Resulaj, R. Kiani, D. M. Wolpert, M. N. Shadlen, Changes of mind in decision-making. *Nature* **461**, 263–266 (2009).
67. P. J. Allen, O. Josephs, R. Turner, A method for removing imaging artifact from continuous EEG recorded during functional MRI. *Neuroimage* **12**, 230–239 (2000).
68. S. Makeig, T. P. Jung, A. J. Bell, T. J. Sejnowski, "Independent component analysis of electroencephalographic data" in *Advances in Neural Information Processing Systems*, D. Touretzky, M. Mozer, M. Hasselmo, Eds. (MIT Press, 1996), vol. 8, pp. 145–151.
69. A. Delorme, J. Palmer, J. Onton, R. Oostenveld, S. Makeig, Independent EEG sources are dipolar. *PLoS One* **7**, e30135 (2012).
70. H. Wickham, The Tidyverse (R Package Version 1.1, 1, 2017).
71. J. Ashburner, A fast diffeomorphic image registration algorithm. *Neuroimage* **38**, 95–113 (2007).
72. N. Tzourio-Mazoyer *et al*., Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* **15**, 273–289 (2002).