nature neuroscience

ARTICLES

https://doi.org/10.1038/s41593-018-0104-6

Neural mediators of changes of mind about perceptual decisions

Stephen M. Fleming[1,2]*, Elisabeth J. van der Putten[3] and Nathaniel D. Daw[4]

Changing one's mind on the basis of new evidence is a hallmark of cognitive flexibility. To revise our confidence in a previous decision, we should use new evidence to update beliefs about choice accuracy. How this process unfolds in the human brain, however, remains unknown. Here we manipulated whether additional sensory evidence supports or negates a previous motion direction discrimination judgment while recording markers of neural activity in the human brain using fMRI. A signature of post-decision evidence (change in log-odds correct) was selectively observed in the activity of posterior medial frontal cortex. In contrast, distinct activity profiles in anterior prefrontal cortex mediated the impact of post-decision evidence on subjective confidence, independently of changes in decision value. Together our findings reveal candidate neural mediators of post-decisional changes of mind in the human brain and indicate possible targets for ameliorating deficits in cognitive flexibility.

John Maynard Keynes allegedly said, "When the facts change, I change my mind." Updating beliefs on the receipt of new evidence is a hallmark of cognitive flexibility. Previous work has focused on how newly arriving evidence for each choice option is evaluated to guide ongoing motor actions in the coordinate frame of a perceptual discrimination decision (for example, left vs. right)[1–4]. However, revising one's confidence about an already-made choice imposes a different coordinate frame on the evidence and requires weighting the evidence comparatively with respect to the choice[5–7]. Here, we devised an extension of a classic motion discrimination task to investigate the computational signatures of such assessment and to investigate how new evidence leads to changes in decision confidence (Fig. 1) while recording markers of neural activity in the human brain using functional magnetic resonance imaging (fMRI). We confirmed behaviorally that post-decision motion led to systematic changes in confidence about the accuracy of a previous decision. This design allowed us to study the underpinnings of changes of mind by analyzing how new evidence impacts confidence bidirectionally, in a graded fashion, rather than only on a subset of trials on which discrete choice reversals are observed.

We hypothesized that brain regions in the human frontal lobe implicated in performance monitoring (posterior medial frontal cortex (pMFC), encompassing dorsal anterior cingulate cortex[8,9] and pre-supplementary motor area[10]) and metacognition (anterior prefrontal cortex; aPFC[11–14]) would play a central role in updating beliefs about previous choice accuracy. Tracking evidence in the coordinate frame of choice accuracy rests on computing a probability that a previous choice was correct or incorrect given the new evidence available, or a change in log-odds correct[5]. When this quantity (which we refer to as 'post-decision evidence' or PDE) is sufficiently low the alternative option becomes more favorable[3]. A Bayesian observer predicts a qualitative signature of PDE in both behavior and neural activity. Specifically, we expect a positive relationship between PDE and motion strength on correct trials (because new evidence serves to a confirm a previous choice) and a negative relationship on error trials (because new evidence disconfirms a previous choice; Fig. 1c, middle).

A further step in the computational chain is to use PDE to update one's final (subjective) confidence in a choice (Fig. 1c, right). For an ideal observer, there is a systematic and direct relationship between PDE and subsequent changes in confidence. However it is known that subjective confidence estimates do not always track objective changes in performance[15,16], and previous studies suggest the prefrontal cortex as a key determinant of such metacognitive fidelity[11,13]. Moreover, a key challenge when interpreting confidence-related neural activity is dissociating distinct variables that may be correlated as a result of a particular task manipulation[17]. For instance, changes in confidence are often correlated with both evidence strength and the expected value of a choice (although see refs. [18,19]). Here we carefully separated these quantities through use of an incentive scheme in which subjects were rewarded for being either highly confident and right or highly unconfident and wrong, ensuring changes in final confidence were decoupled from subjective value (Fig. 1c, right). We also used mediation analyses to formally identify brain activity capturing the impact of model PDE on subjective confidence reports, which were obtained at the end of every trial[20]. This approach has proven fruitful in studying the neural basis of other subjective states, such as pain, while controlling for lower-level effects of sensory stimulation[21], but has not previously been applied to studies of decision-making. Together our findings reveal a division of labor in which pMFC activity tracks post-decision evidence, whereas lateral aPFC also mediates the impact of post-decision evidence on confidence, independently of decision value.

Results

Participants carried out the perceptual decision task outlined in Fig. 1a, first in a behavioral session ($N=25$ subjects) and subsequently while undergoing fMRI ($N=22$ subjects). The subject's goal was to make accurate decisions about the direction of random dot motion and then to estimate confidence in the initial choice. A new sample of dot motion in the same (correct) direction was displayed after the subject's choice but before their confidence rating. Subjects were rewarded for the accuracy of their confidence judgments, and thus the value of a trial increased both when they became more accurate

[1]Wellcome Centre for Human Neuroimaging, University College London, London, UK. [2]Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London, London, UK. [3]Amsterdam Brain and Cognition Center, University of Amsterdam, Amsterdam, The Netherlands. [4]Princeton Neuroscience Institute and Department of Psychology, Princeton University, Princeton, NJ, USA. *e-mail: stephen.fleming@ucl.ac.uk

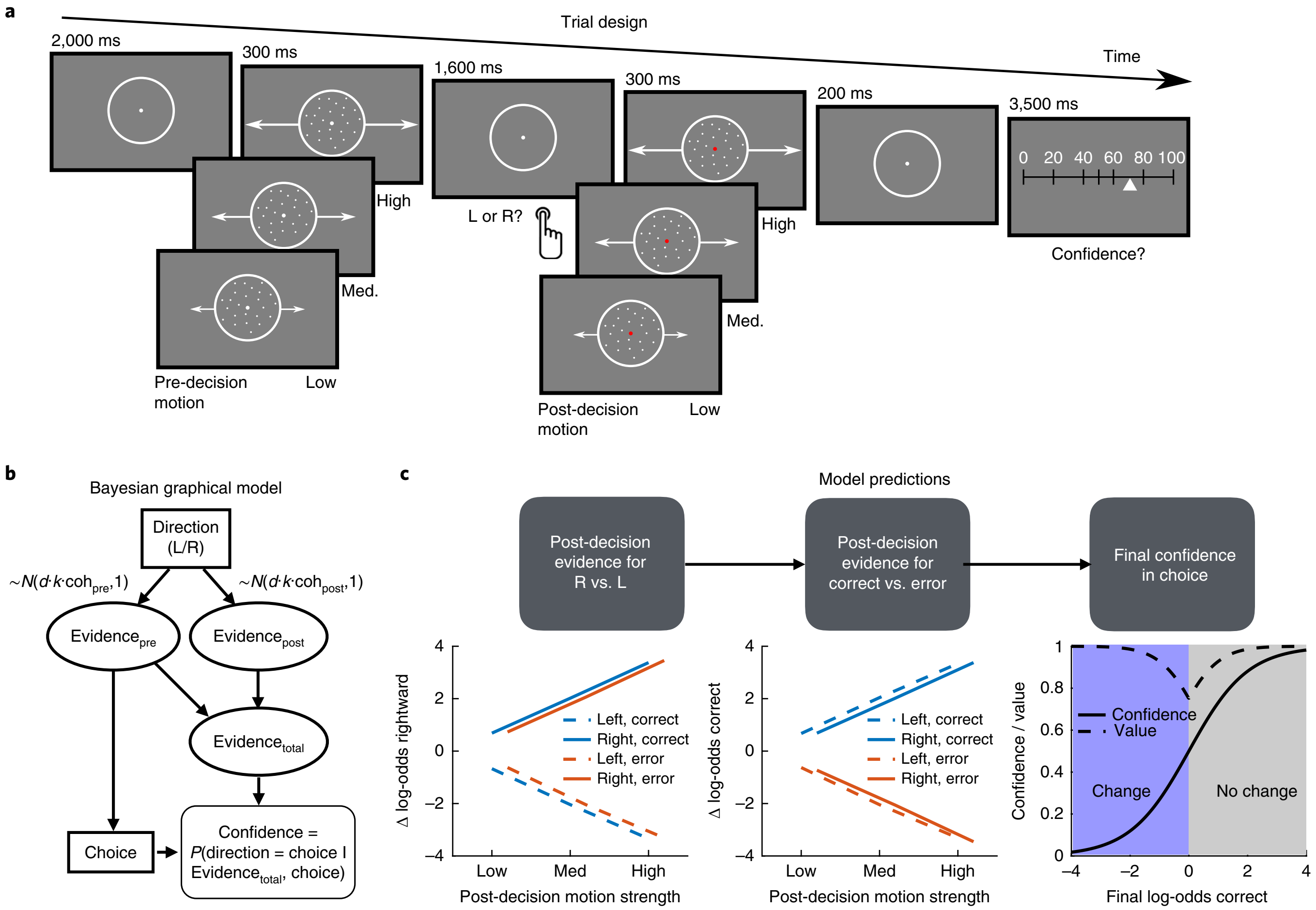

Fig. 1 | Post-decision evidence task and computational framework. a, Task design. Participants made an initial left/right motion discrimination judgment, after which they saw more post-decision motion of variable coherence moving in the same direction as the pre-decision motion. They were asked to rate their confidence in their initial choice on a scale from 0% (certainly wrong) to 100% (certainly correct). Confidence scale steps were further labeled with the words "certainly wrong," "probably wrong," "maybe wrong," "maybe correct," "probably correct" and "certainly correct" (not shown). **b**, Bayesian graphical model indicating how pre- and post-decision motion samples are combined with the chosen action to update an estimate of decision confidence. **c**, Simulated decision variables from the model in **b** showing a distinction between updating evidence in the coordinate frame of motion direction (left) and choice accuracy (middle) as a function of post-decision motion strength and choice. A change in log-odds correct ('post-decision evidence'; PDE) is revealed by a qualitative interaction between post-decision motion strength and choice accuracy (middle). The right panel indicates the expected mapping between log-odds correct and both final confidence and decision value. Confidence and value are dissociated on change-of-mind trials (confidence < 0.5) through use of a quadratic scoring rule that rewards subjects for being either confident and right or unconfident and wrong.

about being right and when they became more accurate about being wrong (see Fig. 1c and Methods). A fully factorial design crossed three pre-decision coherence levels with three post-decision coherence levels, yielding nine experimental conditions. Together these features of the task design allowed us to dissociate motion strength and decision value from changes of mind (Fig. 1c). To equate evidence strength across individuals, before the main task each participant performed a calibration procedure to identify a set of motion coherences that led to approximately 60%, 75% and 90% accuracy (Supplementary Fig. 1). Examination of the empirical cross-correlation between task features and behavior (motion strength, confidence, value and response times) confirmed a limited correlation between predictors (maximum absolute mean $r = 0.38$ for fMRI session; Supplementary Fig. 2).

Choice, confidence and changes of mind. As expected, stronger pre-decision motion led to increases in response accuracy (behavioral session: hierarchical logistic regression, $\beta = 9.21$ (standard error: 0.74), $z = 12.4$, $P < 2 \times 10^{-16}$; fMRI session: $\beta = 7.00$ (0.70), $z = 10.0$, $P < 2.0 \times 10^{-16}$; Fig. 2a,c and Supplementary Table 1). We observed robust changes of confidence in response to post-decision motion (Fig. 2b,d). Specifically, we found that after an erroneous decision, stronger post-decision motion led to progressively lower confidence (behavioral session: hierarchical linear regression, $\beta = -1.15$ (0.14), $\chi^2(1) = 71.8$, $P < 2.2 \times 10^{-16}$; fMRI session: $\beta = -1.05$ (0.11), $\chi^2(1) = 88.0$, $P < 2.2 \times 10^{-16}$; Supplementary Table 2) whereas after a correct decision, confidence was increased as a result of the confirmatory influence of new evidence (behavioral session: $\beta = 0.41$ (0.08), $\chi^2(1) = 26.3$, $P = 3.0 \times 10^{-7}$; fMRI session: $\beta = 0.54$ (0.08), $\chi^2(1) = 44.7$, $P = 2.3 \times 10^{-11}$). Binary changes of mind are revealed by confidence levels lower than 0.5 (i.e., greater confidence in the alternative response), with strong post-decision motion accordingly leading to more frequent binary changes of mind (behavioral session, mean = 11.7% of trials; fMRI session, mean = 18.4% of trials) than weak post-decision motion (behavioral session, mean = 10.4% of trials; fMRI session, mean = 14.8% of trials). Subjects were well calibrated, with final confidence approximately tracking aggregate performance (Supplementary Fig. 3).

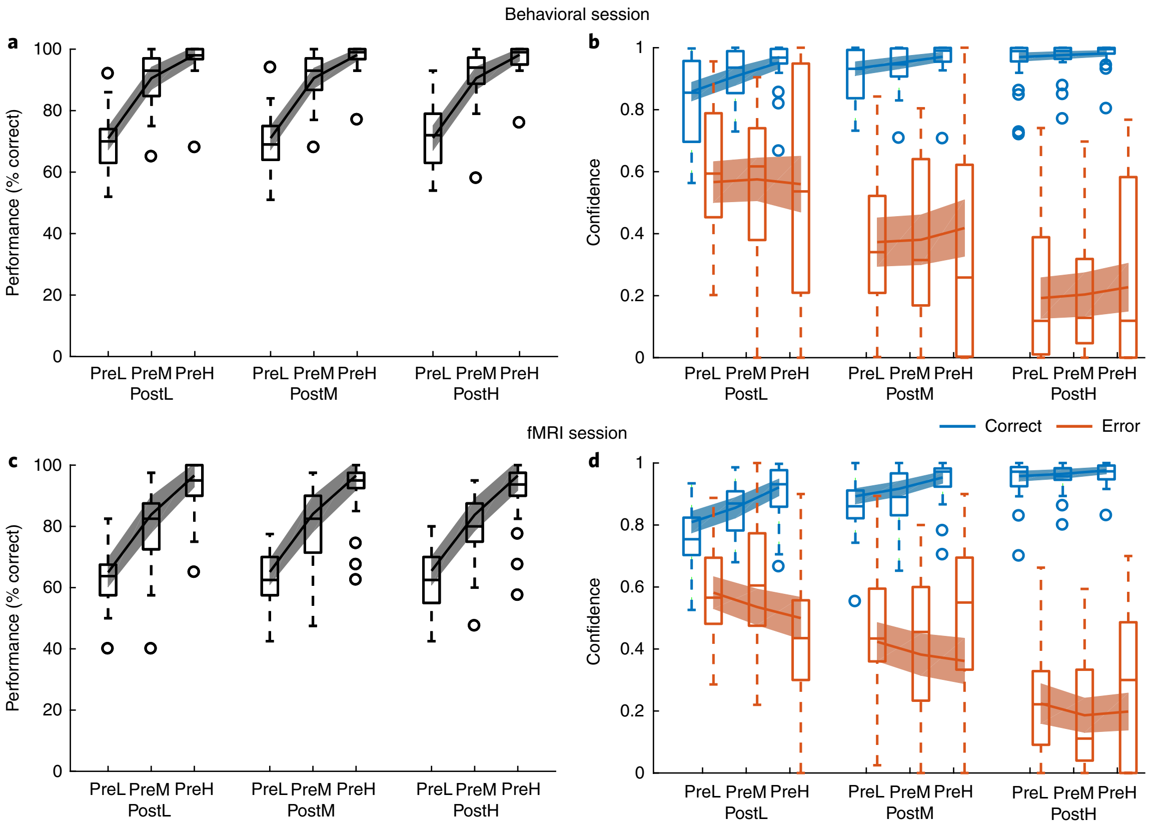

Fig. 2 | Behavioral results. Top, data collected in an initial behavioral session (900 trials per subject, $N = 25$ subjects); bottom, behavioral data collected during the fMRI session (360 trials per subject, $N = 22$ subjects). In each panel data are separated by pre- and post-decision motion coherence (L, low; M, medium; H, high). **a**,**c**, Performance (% correct). **b**,**d**, Aggregate confidence ratings separated according to whether the decision was correct (blue) or incorrect (orange). Lines show data simulated from the best-fitting Bayesian + RT model parameters. Data are plotted as box plots for each condition in which horizontal lines indicate median values, boxes indicate 25–75% interquartile range and whiskers indicate minimum and maximum values; data points outside of 1.5 × the interquartile range are shown separately as circles. For model simulations, error bars reflect 95% confidence intervals for the mean. See also Supplementary Fig. 5.

Computational model of post-decisional change in confidence. We compared a set of alternative computational models of how confidence is affected by post-decision motion strength (see Methods for details). All models generalize signal detection theory, with a single free parameter k mapping pre- and post-decision motion strength (coherence) onto an internal decision variable (Fig. 1b). Extensions to an ideal observer model explored the impact of asymmetric weighting parameters on pre- and post-decision motion[6,7], asymmetric weighting of confirmatory and disconfirmatory evidence[6], flexible mappings between probability correct and reported confidence[22], and the influence of initial response time (RT)[23] (see Methods and Supplementary Fig. 4). We assessed model fit by examining generalization across testing sessions to avoid overfitting; the best-fitting Bayesian + RT model was able to capture both the relationship between pre-decision motion strength and choice accuracy, and the impact of post-decision motion on changes in confidence (Fig. 2 and Supplementary Fig. 5) (difference in median log-likelihood relative to next best model: behavioral→fMRI, 1,932; fMRI→behavioral, 1,298; Supplementary Fig. 4). The β_{RT} parameter of this model was negative in both cases (behavioral session: $\beta_{RT} = -0.73$ (0.26); fMRI session: $\beta_{RT} = -0.37$ (0.22); Supplementary Table 3), indicating that faster initial decisions boosted final confidence. We note that a qualitative signature of PDE in Fig. 1c is common to all model variants and makes clear predictions for interrogation of brain imaging data, to which we turn next.

Neural representations of post-decision evidence. We sought to identify fMRI activity patterns consistent with tracking PDE in the coordinate frame of choice accuracy (changes in log-odds correct due to post-decision motion). Such patterns are characterized by a change in the sign of the relationship between post-decision motion strength and brain activity on correct vs. error trials (Fig. 1c, middle). This change in sign is qualitative, and we remain agnostic about its direction at the level of the fMRI signal: it is plausible that a particular neural population encodes increasing rather than decreasing likelihood of change of mind, in which case we would observe a positive relationship on error trials and a negative relationship on correct trials.

We first computed interaction contrasts (positive or negative) between post-decision motion strength and choice accuracy, to identify patterns of activity that mirror a signature of PDE. Interaction effects were observed whole-brain corrected at both the voxel and cluster level in pMFC (Fig. 3a; peak Montreal Neurological Institute (MNI) coordinates [6 18 50], $P_{\text{voxelFWE}} = 0.002$; $P_{\text{clusterFWE}} < 0.001$) and at the cluster level in right insula (peak [44 14 −6], $P_{\text{clusterFWE}} = 0.009$; Supplementary Table 4). Accordingly, in an independently defined pMFC region of interest (ROI), we obtained a significant interaction between post-decision motion strength and initial decision accuracy in single-trial activity estimates aligned to the onset of post-decision motion (Fig. 3b and Supplementary Table 5; $\beta = -0.11$ (0.037), $\chi^2(1) = 9.35$, $P = 0.0022$). This interaction effect was driven by an increase on error trials and decrease on correct trials (Fig. 3b).

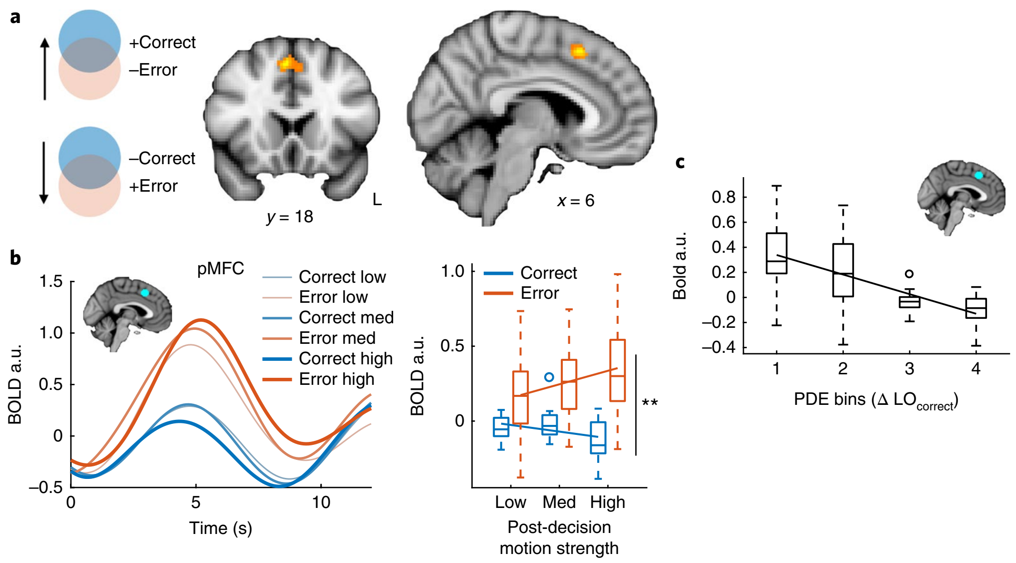

Fig. 3 | Neural signatures of post-decision evidence. a, Whole-brain statistical parametric map for the interaction contrast error/correct × post-decision motion strength, thresholded at $P<0.05$ FWE-corrected, cluster-defining threshold $P<0.001$ (coronal section, $y=18$; sagittal section, $x=6$). Activation in pMFC was significant corrected for multiple comparisons at both the voxel and cluster level (peak MNI coordinate: [6 18 50]). $N=22$ subjects. **b**, fMRI signal extracted from an independent pMFC ROI and sorted according to the subject's choice accuracy (orange, error; blue, correct) and post-decision motion strength. The left panel shows activity time courses aligned to the onset of pre-decision motion (trial start); the right panel shows condition-specific activity estimated from regressors aligned to the onset of post-decision motion. A significant interaction between choice accuracy and post-decision motion strength was obtained in pMFC; hierarchical regression, two-tailed type III Wald χ^2 test, $^{**}P=0.0022$. $N=22$ subjects. **c**, Average blood oxygen level-dependent (BOLD) signal in the pMFC ROI as a function of post-decision evidence extracted from the Bayesian + RT model fit (change in log-odds correct). For visualization, post-decision evidence is collapsed into four equally spaced bins per subject. $N=22$ subjects. In **b** and **c**, BOLD data are plotted as box plots for each condition in which horizontal lines indicate median values, boxes indicate 25–75% interquartile range and whiskers indicate minimum and maximum values; data points outside of 1.5 × the interquartile range are shown separately as circles. Solid lines show the mean of subject-level linear fits.

Finally, to corroborate our model-free analysis, we extracted the predicted PDE ($\mathrm{LO}^{\mathrm{post}}_{\mathrm{correct}}$, where LO is the log posterior odds) on each trial from the Bayesian + RT model fitted to each subject's in-scanner behavioral data. As expected from the model-free pattern, a negative linear relationship was observed between model PDE and pMFC activity (Fig. 3c; $\beta=-0.052$ (0.0085), $\chi^2(1)=37.4$, $P=9.54\times10^{-10}$). No relationship was observed between pre-decision evidence ($\mathrm{LO}^{\mathrm{pre}}_{\mathrm{correct}}$) and pMFC activity ($\beta=-0.013$ (0.013), $\chi^2(1)=1.06$, $P=0.30$), indicating specific engagement during post-decisional changes of confidence. To establish the anatomical specificity of the effect of PDE on brain activity, we interrogated prefrontal and striatal ROIs also implicated in decision confidence and metacognition (ventral striatum, ventromedial prefrontal cortex (vmPFC) and bilateral aPFC areas 46, FPl and FPm from the atlas of Neubert et al.[24]; Supplementary Figs. 6 and 7 and Supplementary Table 5). None of these ROIs showed an interaction between post-decision motion strength and choice ($P>0.05$), and contrasts of regression coefficients revealed greater interaction effects in pMFC compared to aPFC subregions (area 46: $\chi^2(1)=3.7$, $P=0.054$; FPl: $\chi^2(1)=5.0$, $P=0.026$; FPm: $\chi^2(1)=10.9$, $P=0.00095$).

Neural mediators of final confidence. Having identified a putative neural signature of PDE in pMFC, we next searched for brain areas tracking subjects' final confidence in a decision. One computationally plausible hypothesis is that such updates of final confidence are mediated by anatomically distinct networks involved in metacognition[25,26]. aPFC is a leading candidate, as this region is implicated in metacognitive assessment of both perceptual and economic decisions[12,14,18]. In a whole-brain analysis, we found widespread activity showing both positive and negative relationships with final confidence (Fig. 4a and Supplementary Table 6) in regions including pMFC (negative relationship), medial aPFC (positive relationship) and lateral aPFC (negative relationship), consistent with previous studies[12,14,18,27].

We further sought to establish whether aPFC activation continues to track confidence shifts on trials in which discrete changes of mind were recorded (confidence <0.5). Activity that tracks such changes of mind should show a consistent positive or negative slope across both change and no-change trials; in contrast, activity tracking decision value should reverse its relationship with confidence on change trials (owing to the greater reward available for betting against one's choice; Fig. 4b). In a split regression analysis, we found that regression coefficients in lateral aPFC ROIs were significantly negative on both change and no-change trials (Fig. 4b and Supplementary Table 7; area 46: change trials $\beta=-0.36$ (0.15), $\chi^2(1)=5.9$, $P=0.015$; no-change trials $\beta=-0.25$ (0.05), $\chi^2(1)=20.6$, $P=5.6\times10^{-6}$; FPl: change trials $\beta=-0.41$ (0.17), $\chi^2(1)=6.1$, $P=0.013$; no-change trials $\beta=-0.12$ (0.06), $\chi^2(1)=4.1$, $P=0.044$). In contrast, regression coefficients in FPm flipped in sign on change vs. no-change trials (Fig. 4b). Accordingly, when regressing regional time series against both confidence and value in the same general linear model (GLM), we found that confidence but not value covaried with a late signal in area 46 and FPl (Fig. 4c). Conversely, and consistent with previous reports[18,19], FPm (and also pMFC, vmPFC and ventral striatal ROIs; see Supplementary Fig. 7) showed simultaneous correlates of both confidence and value. These results support a conclusion that lateral aPFC subregions are specifically engaged when subjects change their minds about a previous decision on the basis of new evidence.

A key question is how PDE, encoded in pMFC, leads to subsequent shifts in final confidence in a previous decision. To test this hypothesis, we used multi-level mediation analysis[21,28] to jointly test for effects of PDE (from subject-specific fits of the Bayesian + RT

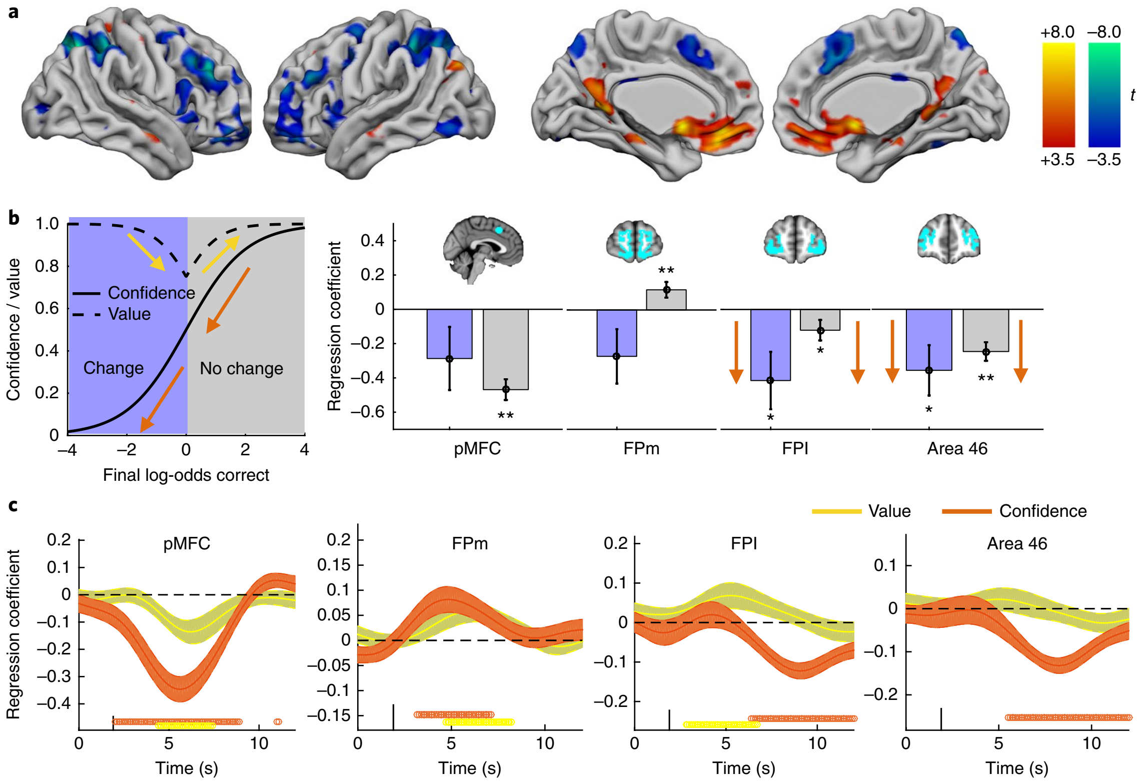

Fig. 4 | Neural signatures of final confidence in choice. a, Whole-brain analysis of activity related to final confidence reports on each trial. Cool colors indicate negative relationships; hot colors indicate positive relationships. Thresholded at $P < 0.05$, FWE-corrected for multiple comparisons, cluster-defining threshold $P < 0.001$. $N = 22$ subjects. **b**, Hierarchical regression coefficients relating confidence to single-trial activity estimates on both change-of-mind and no-change-of-mind trials. Orange arrows indicate that the pattern of coefficients is consistent in sign, as predicted for regions tracking the full range of final confidence in an initial choice. Yellow arrows indicate a flip in sign, as predicted for regions tracking changes in decision value. Error bars indicate standard errors of the coefficient means. $^{**}P < 0.01$, $^{*}P < 0.05$, two-tailed type III Wald χ^2 test; see Supplementary Table 7. $N = 22$ subjects. **c**, Multiple regressions of confidence and value on activity time course in ROIs. Points below time course indicate significant excursions of *t*-statistics assessed using two-tailed permutation tests. Error bars indicate standard errors of the coefficient mean. $N = 22$ subjects.

computational model) on brain activity (path *a*), brain activity on final confidence (path *b*) and mediation ($a \times b$) effects (Fig. 5), while controlling for both response time and pre-decision evidence. A mediator can be interpreted as an indirect pathway through a brain area that links PDE with changes in subjective confidence, suggesting that if such a region were disrupted, this relationship would also be disrupted or abolished. We examined mediation both in anatomically defined aPFC subregions and at the voxel level across the whole brain.

In line with our hypothesis, activity in area 46 and FPl was found to mediate the impact of PDE on final confidence (Fig. 5a and Supplementary Table 8; $a \times b$ effect, bootstrapped *P*-values: area 46, $P = 0.0027$; FPl, $P = 0.0056$). While mediation modeling is correlational, precluding a direct inference as to directionality, we note that control models in which PDE and confidence were reversed did not result in a significant mediation effect in either area 46 ($P = 0.54$) or FPl ($P = 0.46$). Mediation may be driven either by consistent effects of paths *a* and *b* across the group or by covariance between stimulus- and report-related responses[21]. In area 46 there was evidence for consistent main effects of path *a* and *b* in the group as a whole. In contrast, in FPl, mediation was driven by the covariance of *a* and *b* paths across subjects. Finally, in a voxel-based mediation analysis, we observed a significant cluster in left lateral aPFC (Fig. 5b), corroborating our ROI analysis.

In an exploratory whole-brain analysis we also observed clusters in pMFC and bilateral parietal cortex that, together with aPFC, met whole-brain corrected statistical criteria for mediation (Supplementary Fig. 8). This result is consistent with pMFC activity both tracking PDE (Fig. 3c) and covarying with final confidence (Fig. 4a). Taken together, our findings indicate complementary roles for frontal subregions in changes of mind: pMFC (but not aPFC) activity tracks PDE, whereas lateral aPFC also mediates changes in final confidence estimates, independently of decision value.

Discussion

Changing one's mind on the basis of new evidence is a hallmark of cognitive flexibility. Such reversals are supported computationally by sensitivity to post-decision evidence: if I have made an error and the new evidence is compelling, I should change my mind. Here we devised a manipulation of post-decisional information in perceptual decision-making to study this process. Participants appropriately increased their confidence when new evidence was supportive of an initial decision and decreased their confidence when it was contradictory. A signature of post-decision evidence encoding, a change in log-odds correct, was identified in the activity of pMFC. We further observed that distinct activity profiles in lateral aPFC mediated the impact of post-decision evidence on subjective confidence.

Previous work has focused on how stimulus evidence may reverse the accumulation of evidence in circuits coding for one or the other choice option (for example, left or right). To update one's confidence in a previous choice, new evidence in the coordinate frame of stimulus/response may be further transformed into the coordinate frame of choice accuracy[5]. These schemes are not mutually

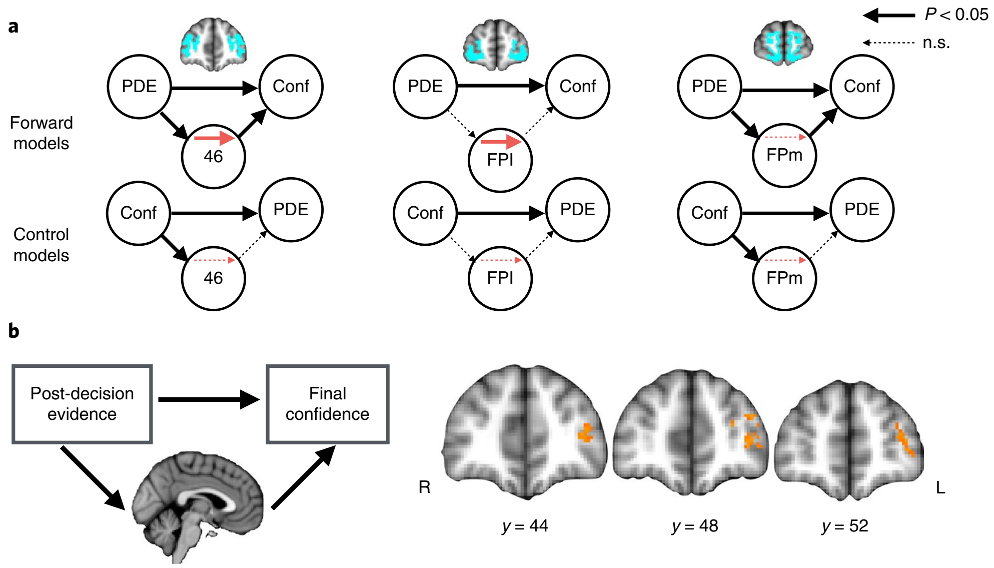

Fig. 5 | Neural mediation of PDE on final confidence. a, Multi-level mediation analysis assessing whether the effect of PDE on final confidence is mediated by activity in anatomically defined aPFC ROIs. For each ROI, the upper row of models indicates forward mediation and the lower row indicates reverse mediation (of confidence onto PDE). Mediation was observed only for forward models in areas 46 and FPl (red arrows). Arrow thickness reflects two-tailed bootstrapped *P*-values; see Supplementary Table 8 for statistics. $N = 22$ subjects. **b**, The model used in **a** was fit to each voxel independently to create a map of *P*-values for the mediation ($a \times b$) effect in aPFC. Thresholded at $P < 0.05$ FWE-corrected at the cluster level using Monte Carlo simulation, cluster-defining threshold $P < 0.001$. $N = 22$ subjects. See also Supplementary Fig. 8.

exclusive: to update an ongoing action plan, it may be sufficient to continue accumulating evidence in a 'pipeline' directly guiding the movement toward one or the other target[2,3] while in parallel revising one's belief in the accuracy of a previous choice[25,26]. In an elegant behavioral study, van den Berg and colleagues demonstrated that a single stream of evidence may continue to accumulate during action initiation and, via a comparison to thresholds specified in stimulus/response space (e.g., log-odds rightward), be used to guide changes of both decision and (response-specific) confidence[3]. Here, by introducing a manipulation of post-decisional information, we reveal a circumscribed activity pattern in pMFC consistent with tracking PDE in the frame of choice accuracy. Examining mutual interactions between evidence coded in the frame of stimulus/response identity or choice accuracy is beyond the design of the current study, but may be profitably investigated by tracking each of these coordinate frames using techniques with high temporal resolution such as magnetoencephalography.

Even in the absence of a direct manipulation of post-decision evidence, signal detection models of decision confidence predict an interaction between stimulus strength and choice accuracy[16,29]. We also observed such a pattern in our behavioral data: confidence decreased on error trials and increased on correct trials, when pre-decision motion was stronger (Supplementary Table 2; this effect was tempered by the influence of response times on error-trial confidence, as shown by the fits of the Bayesian + RT model in Fig. 2). However, we note that the interaction effect in pMFC was primarily driven by post- and not pre-decision evidence (Supplementary Table 5), indicating a distinct role in post-decisional changes of mind. An interaction between stimulus strength and choice accuracy has also been observed in the activity of rodent orbitofrontal cortex in the absence of a post-decision evidence manipulation[29], and inactivation of this region impairs confidence-guided behaviors[30]. Searching for signatures of PDE in other species may therefore shed light on mechanisms supporting changes of mind that are conserved (for example, in homologs of pMFC[31]) and those that may be unique to humans (for example, those supported by granular aPFC).

The function of pMFC in human cognition has been the subject of extended scrutiny and debate. A well-established finding is that a paracingulate region activates to error commission, consistent with its role as a cortical generator of the error-related negativity[8–10]. More recently, studies have linked dorsal anterior cingulate activity to a broader role in behavioral switching away from a default option[32]. Our findings complement these lines of work by characterizing a computation related to changes of mind. Specifically, our analysis indicates that pMFC activity tracks whether an initial choice should be revised in light of newly acquired information. The peak activation in this contrast was obtained in pre-supplementary motor area, dorsal to the rostral cingulate zone[33]. While previous studies of error detection have focused on all-or-nothing, endogenous error responses in pMFC, our findings suggest a more computationally sophisticated picture: pMFC activity tracked graded changes in log-odds correct[34,35] (Fig. 3c). Together our results indicate that error monitoring, confidence and changes of mind may represent different behavioral manifestations of a common computation supported by inputs to pMFC[25,36,37].

Beyond pMFC, we found a widespread network of regions where activity tracks final confidence, including negative correlations in lateral PFC, parietal cortex and pMFC, and positive correlations in vmPFC and precuneus, consistent with previous findings[12,14,18,19,27]. Building on an analogous body of work on the neural substrates of subjective pain[21,38], we used mediation analysis to formally disentangle the inter-relationships among post-decision evidence, brain activity and the final confidence subjects held in their decision. Lateral aPFC (areas 46 and FPl) activity mediated the impact of post-decision evidence on subjective confidence. Lateral aPFC has previously been implicated in self-evaluation of decision performance[12,14,18], and it receives an anatomical projection from pMFC[39]. It is notable that in the current study the activity profile of lateral aPFC covaried with final confidence in both mediation and regression analyses, but did not track post-decision evidence or decision value per se. It is therefore plausible that lateral aPFC supports a representation of choice quality that contributes to metacognitive control of future behavior[40–42]. Together with aPFC, posterior parietal cortex was also implicated by exploratory whole-brain analyses as a mediator of the impact of PDE on confidence, consistent with a role for a broader frontoparietal network in metacognition and confidence formation[43,44].

In previous research it has proven difficult to isolate changes in decision confidence from other confounding variables. The probability of a previous decision remaining correct is often correlated with expected value. In other words, if subjects are motivated to be accurate, decision confidence usually scales with decision value. Here we separated expected value from confidence by allowing subjects to gain rewards by betting against their original decision using the quadratic scoring rule. This rule returns maximum reward both when a correct trial is rated with high confidence and when an incorrect trial is rated with low confidence (Fig. 1c). In medial PFC we found a U-shaped pattern of activity in relation to reported confidence, consistent with previous findings that both confidence and value are multiplexed on the medial surface[18,19]. In contrast, lateral aPFC activity covaried with final confidence reports but not value, indicating a specific role in changes of mind.

In conclusion, by integrating computational modeling with human fMRI, we reveal a neural signature of how new evidence is integrated to support graded changes of mind. Multiple coordinate frames are in play when new evidence leads to shifts in beliefs—from coding evidence in support of one or the other decision option, to updating the accuracy of a choice, to communicating changes in confidence. Neuroimaging revealed complementary roles for frontal subregions in changes of mind: post-decision evidence was tracked by pMFC while aPFC also mediated final confidence in choice. Failure of such updating processes may lead to impairments to cognitive flexibility and/or an inability to discard previously held beliefs[45,46]. Together our findings shed light on the building blocks of changes of mind in the human brain and indicate possible targets for amelioration of such deficits.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41593-018-0104-6.

Received: 4 August 2017; Accepted: 26 January 2018;
Published online: 12 March 2018

References

1. Kiani, R., Cueva, C. J., Reppas, J. B. & Newsome, W. T. Dynamics of neural population responses in prefrontal cortex indicate changes of mind on single trials. *Curr. Biol.* **24**, 1542–1547 (2014).
2. Resulaj, A., Kiani, R., Wolpert, D. M. & Shadlen, M. N. Changes of mind in decision-making. *Nature* **461**, 263–266 (2009).
3. van den Berg, R. et al. A common mechanism underlies changes of mind about decisions and confidence. *Elife* **5**, e12192 (2016).
4. Pleskac, T. J. & Busemeyer, J. R. Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychol. Rev.* **117**, 864–901 (2010).
5. Pouget, A., Drugowitsch, J. & Kepecs, A. Confidence and certainty: distinct probabilistic quantities for different goals. *Nat. Neurosci.* **19**, 366–374 (2016).
6. Bronfman, Z. Z. et al. Decisions reduce sensitivity to subsequent information. *Proc. Biol. Sci.* **282**, 20150228 (2015).
7. Yu, S., Pleskac, T. J. & Zeigenfuse, M. D. Dynamics of postdecisional processing of confidence. *J. Exp. Psychol. Gen.* **144**, 489–510 (2015).
8. Carter, C. S. et al. Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science* **280**, 747–749 (1998).
9. Dehaene, S., Posner, M. I. & Tucker, D. M. Localization of a neural system for error detection and compensation. *Psychol. Sci.* **5**, 303–305 (1994).
10. Bonini, F. et al. Action monitoring and medial frontal cortex: leading role of supplementary motor area. *Science* **343**, 888–891 (2014).
11. Fleming, S. M., Ryu, J., Golfinos, J. G. & Blackmon, K. E. Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain* **137**, 2811–2822 (2014).
12. Fleming, S. M., Huijgen, J. & Dolan, R. J. Prefrontal contributions to metacognition in perceptual decision making. *J. Neurosci.* **32**, 6117–6125 (2012).
13. Shimamura, A. P. & Squire, L. R. Memory and metamemory: a study of the feeling-of-knowing phenomenon in amnesic patients. *J. Exp. Psychol. Learn. Mem. Cogn.* **12**, 452–460 (1986).
14. Hilgenstock, R., Weiss, T. & Witte, O. W. You'd better think twice: post-decision perceptual confidence. *Neuroimage* **99**, 323–331 (2014).
15. Fleming, S. M. & Lau, H. C. How to measure metacognition. *Front. Hum. Neurosci.* **8**, 443 (2014).
16. Sanders, J. I., Hangya, B. & Kepecs, A. Signatures of a statistical computation in the human sense of confidence. *Neuron* **90**, 499–506 (2016).
17. Rushworth, M. F. S. & Behrens, T. E. J. Choice, uncertainty and value in prefrontal and cingulate cortex. *Nat. Neurosci.* **11**, 389–397 (2008).
18. De Martino, B., Fleming, S. M., Garrett, N. & Dolan, R. J. Confidence in value-based choice. *Nat. Neurosci.* **16**, 105–110 (2013).
19. Lebreton, M., Abitbol, R., Daunizeau, J. & Pessiglione, M. Automatic integration of confidence in the brain valuation signal. *Nat. Neurosci.* **18**, 1159–1167 (2015).
20. Baron, R. M. & Kenny, D. A. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.* **51**, 1173–1182 (1986).
21. Atlas, L. Y., Lindquist, M. A., Bolger, N. & Wager, T. D. Brain mediators of the effects of noxious heat on pain. *Pain* **155**, 1632–1648 (2014).
22. Zhang, H. & Maloney, L. T. Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. *Front. Neurosci.* **6**, 1 (2012).
23. Kiani, R., Corthell, L. & Shadlen, M. N. Choice certainty is informed by both evidence and decision time. *Neuron* **84**, 1329–1342 (2014).
24. Neubert, F.-X., Mars, R. B., Thomas, A. G., Sallet, J. & Rushworth, M. F. S. Comparison of human ventral frontal cortex areas for cognitive control and language with areas in monkey frontal cortex. *Neuron* **81**, 700–713 (2014).
25. Fleming, S. M. & Daw, N. D. Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychol. Rev.* **124**, 91–114 (2017).
26. Insabato, A., Pannunzi, M., Rolls, E. T. & Deco, G. Confidence-related decision making. *J. Neurophysiol.* **104**, 539–547 (2010).
27. Fleck, M. S., Daselaar, S. M., Dobbins, I. G. & Cabeza, R. Role of prefrontal and anterior cingulate regions in decision-making processes shared by memory and nonmemory tasks. *Cereb. Cortex* **16**, 1623–1630 (2006).
28. Kenny, D. A., Korchmaros, J. D. & Bolger, N. Lower level mediation in multilevel models. *Psychol. Methods* **8**, 115–128 (2003).
29. Kepecs, A., Uchida, N., Zariwala, H. A. & Mainen, Z. F. Neural correlates, computation and behavioural impact of decision confidence. *Nature* **455**, 227–231 (2008).
30. Lak, A. et al. Orbitofrontal cortex is required for optimal waiting based on decision confidence. *Neuron* **84**, 190–201 (2014).
31. Wallis, J. D. Cross-species studies of orbitofrontal cortex and value-based decision-making. *Nat. Neurosci.* **15**, 13–19 (2011).
32. Kolling, N., Behrens, T. E. J., Mars, R. B. & Rushworth, M. F. S. Neural mechanisms of foraging. *Science* **336**, 95–98 (2012).
33. Neubert, F.-X., Mars, R. B., Sallet, J. & Rushworth, M. F. S. Connectivity reveals relationship of brain areas for reward-guided learning and decision making in human and monkey frontal cortex. *Proc. Natl Acad. Sci. USA* **112**, E2695–E2704 (2015).
34. Boldt, A. & Yeung, N. Shared neural markers of decision confidence and error detection. *J. Neurosci.* **35**, 3478–3484 (2015).
35. Scheffers, M. K. & Coles, M. G. H. Performance monitoring in a confusing world: error-related brain activity, judgments of response accuracy, and types of errors. *J. Exp. Psychol. Hum. Percept. Perform.* **26**, 141–151 (2000).
36. Yeung, N. & Summerfield, C. Metacognition in human decision-making: confidence and error monitoring. *Phil. Trans. R. Soc. B* **367**, 1310–1321 (2012).
37. Murphy, P. R., Robertson, I. H., Harty, S. & O'Connell, R. G. Neural evidence accumulation persists after choice to inform metacognitive judgments. *Elife* **4**, 3478 (2015).
38. Atlas, L. Y., Bolger, N., Lindquist, M. A. & Wager, T. D. Brain mediators of predictive cue effects on perceived pain. *J. Neurosci.* **30**, 12964–12977 (2010).
39. Liu, H. et al. Connectivity-based parcellation of the human frontal pole with diffusion tensor imaging. *J. Neurosci.* **33**, 6782–6790 (2013).
40. Badre, D., Doll, B. B., Long, N. M. & Frank, M. J. Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration. *Neuron* **73**, 595–607 (2012).
41. Purcell, B. A. & Kiani, R. Hierarchical decision processes that operate over distinct timescales underlie choice and changes in strategy. *Proc. Natl Acad. Sci. USA* **113**, E4531–E4540 (2016).
42. Shea, N. et al. Supra-personal cognitive control and metacognition. *Trends Cogn. Sci.* **18**, 186–193 (2014).

43. Cortese, A., Amano, K., Koizumi, A., Kawato, M. & Lau, H. Multivoxel neurofeedback selectively modulates confidence without changing perceptual performance. *Nat. Commun.* **7**, 13669 (2016).
44. Kiani, R. & Shadlen, M. N. Representation of confidence associated with a decision by neurons in the parietal cortex. *Science* **324**, 759–764 (2009).
45. Moritz, S. & Woodward, T. S. A generalized bias against disconfirmatory evidence in schizophrenia. *Psychiatry Res.* **142**, 157–165 (2006).
46. Woodward, T. S., Buchy, L., Moritz, S. & Liotti, M. A bias against disconfirmatory evidence is associated with delusion proneness in a nonclinical sample. *Schizophr. Bull.* **33**, 1023–1028 (2007).

Acknowledgements

We thank D. Bang and B. De Martino for comments on an earlier draft of this manuscript. Funded by a Sir Henry Wellcome Fellowship from the Wellcome Trust (WT096185) awarded to S.M.F. N.D.D is funded by the James S. McDonnell Foundation and the John Templeton Foundation. The Wellcome Centre for Human Neuroimaging is supported by core funding from the Wellcome Trust (203147/Z/16/Z).

Author contributions

S.M.F. designed experiments, performed experiments, analyzed behavioral and neuroimaging data, developed computational models and wrote the paper; E.J.v.d.P. performed experiments and analyzed behavioral data; N.D.D. designed experiments, developed computational models and wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41593-018-0104-6.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to S.M.F.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Participants. Twenty-five participants gave written informed consent to take part in a study conducted across two separate days. No statistical tests were used to predetermine the sample size, which is similar to those reported in previous publications[14,32,40]. A behavioral experiment was administered on the first day and an fMRI experiment on the second day. Twenty-five participants were included in the analysis of behavioral data (14 females, mean age 24.0, s.d. = 3.6). In the fMRI experiment, one participant was excluded because of excess head motion and one participant was excluded because of lack of variability in confidence ratings (308/360 trials were rated as 100% confident). A further participant attended only the first behavioral session. Twenty-two participants were included in the analysis of fMRI data (12 females, mean age 24.1, s.d. = 3.4). The study was approved by NYU's University Committee on Activities Involving Human Subjects, all relevant ethical regulations were followed, and participants provided written consent before the experiment.

Stimuli. The experiment was programmed in Matlab 2014b (MathWorks) using Psychtoolbox (version 3.0.12)[47,48]. In the behavioral session, stimuli were presented on an iMac desktop monitor viewed at a distance of approximately 45 cm. In the scanner, stimuli were presented via a projector at an approximate viewing distance of 58 cm. Stimuli consisted of random-dot kinematograms (RDKs). Each RDK consisted of a field of random dots (0.12° diameter) contained in a 7° circular white aperture. Each set of dots lasted for one video frame and was replotted three frames later[49]. Each time the same set of dots was replotted, a subset determined by the percent coherence was offset from their original location in the direction of motion and the remaining dots were replotted randomly. Motion direction was either to the left or right along the horizontal meridian. Coherently moving dots moved at a speed of $5° \text{ s}^{-1}$ and the number of dots in each frame was specified to create a density of $30 \text{ dots deg}^{-2} \text{ s}^{-1}$. Each RDK lasted for 300 ms.

Task and procedure. Participants attended the laboratory on two different days. On the first day they completed a calibration session to obtain their psychometric function for motion discrimination, followed by 900 trials of the main experiment shown in Fig. 1a. On the second day participants completed the fMRI scan. Data collection and analysis were not performed blind to the conditions of the experiments.

Behavioral session. *Calibration phase.* Before performing the main task, each participant performed 240 trials of motion direction estimation without confidence ratings or further post-decision motion. These trials were equally distributed across six coherence levels: 3%, 8%, 12%, 24%, 48% and 100%. Motion direction (left or right) was randomized and independent of coherence. Judgments were made using the left or right arrow keys on a standard computer keyboard after the offset of each stimulus, and the response was without time limit. During the calibration phase (but not the experiment phase), auditory feedback was delivered to indicate whether the judgment was correct (high-pitched tone) or incorrect (low-pitched tone). The intertrial interval was 1 s. The three coherence levels that resulted in 60%, 75% and 90% correct choices were individually determined for each subject using probit regression. These coherence levels were then stored for use in the experiment phase.

Experiment phase. In the main experiment, subjects completed 900 trials of the task shown in Fig. 1a. Each trial consisted of the following events in sequence. A central fixation point (0.2° diameter) and empty aperture were presented, followed by an RDK of low, medium or high coherence. Following the offset of the RDK, participants were asked to make a judgment as to whether the movement of the dots was to the left or the right. Their response triggered a second post-decision RDK that was shown after a delay of 100 ms. The second post-decision RDK was always in the same (correct) direction as the first pre-decision RDK, but of a variable coherence. Subjects were instructed that this was "bonus" motion that they could use to inform their confidence in their initial response. They were told that the bonus motion was always in the same direction as the regular motion, but were not informed that it may have varied in strength. A fully factorial design crossed 3 pre-decision coherence levels with 3 post-decision coherence levels, yielding 9 experimental conditions, each with 100 trials. Trial order was fully randomized for each subject.

After the bonus motion was displayed, an empty aperture was presented for 200 ms and then participants were asked to indicate their confidence in their initial judgment on a horizontal scale (length = 14°) ranging from 0 to 100%. Confidence responses were made with a mouse click controlled by the right hand and could be made anywhere along the scale. Half of subjects saw the scale labeled with 0% on the left and 100% on the right and half saw the reverse orientation, with scale orientation fixed across both the behavioral and fMRI sessions. A vertical red cursor provided feedback as to the selected rating. In the behavioral session there was no time limit for either the response or the confidence rating, and no feedback was given as to whether the response was correct or incorrect.

fMRI session. During the structural scan at the start of the fMRI experiment, participants carried out a 'top-up' calibration session consisting of 120 trials of left/right motion judgments without confidence ratings. Three randomly interleaved QUEST adaptive staircases were used to estimate coherence levels associated with 60%, 75% and 90% correct performance. The prior for each staircase was centered on the corresponding coherence estimate derived from the behavioral calibration session.

Before entering the scanner, participants were refamiliarized with the task and confidence rating scale. The task was identical to that described above except for the following changes. Response deadlines of 1.5 s and 3 s were imposed for the initial decision and confidence rating, respectively. Both motion judgments and confidence ratings were made via an fMRI button box held in the right hand. To rate confidence, participants used their index and middle fingers to move a cursor in steps of 10% to the left or right of the scale. The initial cursor location on each trial was randomized. The rating was confirmed by pressing a third button with the ring finger, after which the cursor changed from white to red for 500 ms. During each of the 4 scanner runs participants completed 90 trials.

After the main experiment, we carried out a localizer scan for motion-related activity. During this scan participants passively viewed 20 alternating displays of moving and stationary dots, each lasting 12 s. Equal numbers of leftward and rightward moving dot displays were included at a constant coherence of 50%.

Scoring rule for confidence ratings. Confidence ratings were incentivized using the quadratic scoring rule (QSR)[50]:

$$\text{points} = 100\left[1-(\text{correct}_i - \text{conf}_i)^2\right]$$

where correct_i is equal to 1 on trial i if the choice was correct and 0 otherwise, and conf_i is the subject's confidence rating on trial i entered as a probability between 0 and 1. The QSR is a proper scoring rule in that maximum earnings are obtained by jointly maximizing the accuracy of choices and confidence ratings[51]. For every 5,000 points, subjects received an extra $1. This scoring rule ensures that confidence is orthogonal to the reward the subject expects to receive for each trial. Maximal reward is obtained both when one is maximally confident and right and when one is minimally confident and wrong (Fig. 1c).

The confidence scale was labeled both with scale steps of 0%, 20%, 40%, 60%, 80% and 100% (positioned above the line) and, following Boldt and Yeung[34], verbal confidence labels of "certainly wrong," "probably wrong," "maybe wrong," "maybe correct," "probably correct" and "certainly correct" (positioned below the line). The scale midpoint was marked with a vertical tick halfway between the 40% and 60% labels. Before taking part in the main experiment, participants underwent a training session to instruct them in the use of the confidence scale. Following Moore and Healy[52], participants were first instructed:"You can win points by matching your confidence to your performance. Specifically, the number of points you earn is based on a rule that calculates how closely your confidence tracks your performance: $points = 100^{*}\left[1-(accuracy - confidence)^2\right]$.

This formula may appear complicated, but what it means for you is very simple: You will get paid the most if you honestly report your best guess about the likelihood of being correct. You can earn between 0 and 100 points for each trial."

Participants were then asked where they should click on the scale if they were sure they responded either correctly or incorrectly. They were then informed:"The correct answers were: If you are sure you responded correctly, you should respond 100% confidence/certainly correct. If you are sure you picked the wrong direction, you should respond 0% confidence/certainly wrong. If you are not 100% sure about being correct or incorrect you should select a location in between according to the following descriptions on the confidence scale: probably incorrect = 20% confidence; maybe incorrect = 40% confidence; maybe correct = 60% confidence; probably correct = 80% confidence. You can also click anywhere in between these percentages."

Statistics. Effects of condition on confidence ratings and accuracy were assessed using hierarchical mixed-effects regression using the lme4 package in R (version 3.3.3)[53]. For confidence ratings, we constructed linear models separately for correct and incorrect trials. Pre- and post-decision coherence values and their interaction were entered as separate predictors of confidence. Log response times were also included in the model. We obtained P-values for regression coefficients using the car package for R[54]. Mixed-effects logistic regression was used to quantify the effect of condition on response accuracy. In all regressions we modeled subject-level slopes and intercepts, and report coefficients and statistics at the population level. The distribution of residuals in regression models was assumed to be normal, but this was not formally tested.

Bayesian model. We developed a Bayesian model of choice and confidence that is grounded in signal detection theory. Subjects receive two internal samples, X_{pre} generated from pre-decision motion and X_{post} from post-decision motion. Motion direction $d \in [-1, 1]$ determines the sample means with Gaussian signal-to-noise depending linearly on coherence θ_{pre} or θ_{post} via sensitivity parameter k *(where ~ indicates "is distributed as")*:

$$X_{\text{pre}} \sim N(dk\theta_{\text{pre}}, 1)$$

$$X_{\mathrm{post}} \sim N\left(dk\theta_{\mathrm{post}}, 1\right)$$

We assume that subjects do not know the coherence levels on a particular trial, θ_{pre} and θ_{post}, which are nuisance parameters that do not carry any information about the correct choice. We therefore approximate the likelihood of X_{pre} and X_{post} as a Gaussian with mean μ and variance σ^2 determined by a mixture of Gaussians across each of the three possible coherence levels. Starting with X_{pre}:

$$P\left(X_{\mathrm{pre}} \mid d=1\right)=\sum_{\theta_{\mathrm{pre}}} p\left(\theta_{\mathrm{pre}}\right) N\left(k\theta_{\mathrm{pre}}, 1\right)$$

As each of the three coherence levels are equally likely by design $\left(p\left(\theta_{\mathrm{pre}}\right)=0.33\right)$, we can define the mean as

$$\mu=\frac{\sum k\theta_{\mathrm{pre}}}{3}$$

The aggregate variance σ^2 can be decomposed into both between- and within-condition variance. From the law of total variance:

$$\sigma^2=\sum_{\theta_{\mathrm{pre}}} p\left(\theta_{\mathrm{pre}}\right)\left[E\left[X_{\mathrm{pre}} \middle| k\theta_{\mathrm{pre}}\right]-\mu\right]^2+\sum_{\theta_{\mathrm{pre}}} p\left(\theta_{\mathrm{pre}}\right) \operatorname{Var}\left(X_{\mathrm{pre}} \middle| k\theta_{\mathrm{pre}}\right)$$

$$\sigma^2=\sum_{\theta_{\mathrm{pre}}} p\left(\theta_{\mathrm{pre}}\right)\left[k\theta_{\mathrm{pre}}-\mu\right]^2+1$$

Because the possible values of θ are the same pre- and post-decision, μ and σ^2 are the same for both X_{pre} and X_{post}. Actions a are made by comparing X_{pre} to a criterion parameter m that accommodates any stimulus-independent biases toward the leftward or rightward response, $a=\operatorname{sign}(X_{\mathrm{pre}}-m)$.

Each sample, X_{pre} and X_{post}, updates the log posterior odds of motion direction (rightward or leftward), $\mathrm{LO}_{\mathrm{dir}}$, which under flat priors is equal to the log-likelihood:

$$\mathrm{LO}_{\mathrm{dir}}^{\mathrm{pre}}=log\frac{P(d=1|X_{\mathrm{pre}})}{P(d=-1|X_{\mathrm{pre}})}=log\frac{P(X_{\mathrm{pre}}|d=1)}{P(X_{\mathrm{pre}}|d=-1)}$$

$$\mathrm{LO}_{\mathrm{dir}}^{\mathrm{post}}=log\frac{P(d=1|X_{\mathrm{post}})}{P(d=-1|X_{\mathrm{post}})}=log\frac{P(X_{\mathrm{post}}|d=1)}{P(X_{\mathrm{post}}|d=-1)}$$

where, as a result of the Gaussian generative model for X, $\mathrm{LO}_{\mathrm{dir}}$ is as follows:

$$\mathrm{LO}_{\mathrm{dir}}=log\frac{\mathrm{e}^{(\mu+X)^2/2\sigma^2}}{\mathrm{e}^{(\mu-X)^2/2\sigma^2}}$$

$$\mathrm{LO}_{\mathrm{dir}}=\frac{2\mu X}{\sigma^2}$$

The total accumulated evidence for rightward vs. leftward motion at the end of the trial is

$$\mathrm{LO}_{\mathrm{dir}}^{\mathrm{total}}=\mathrm{LO}_{\mathrm{dir}}^{\mathrm{pre}}+\mathrm{LO}_{\mathrm{dir}}^{\mathrm{post}}$$

Positive values indicate greater belief in rightward motion; negative values, greater belief in leftward motion.

To update confidence in one's choice, the belief in motion direction ($\mathrm{LO}_{\mathrm{dir}}$) is transformed into a belief about decision accuracy ($\mathrm{LO}_{\mathrm{correct}}$) conditional on the chosen action:

If $a=1$:

$$\mathrm{LO}_{\mathrm{correct}}=\mathrm{LO}_{\mathrm{dir}}$$

Otherwise:

$$\mathrm{LO}_{\mathrm{correct}}=-\mathrm{LO}_{\mathrm{dir}}$$

As with $\mathrm{LO}_{\mathrm{dir}}$, $\mathrm{LO}_{\mathrm{correct}}$ can be decomposed into pre- and post-decisional components:

$$\mathrm{LO}_{\mathrm{correct}}^{\mathrm{total}}=\mathrm{LO}_{\mathrm{correct}}^{\mathrm{pre}}+\mathrm{LO}_{\mathrm{correct}}^{\mathrm{post}}$$

The final log odds correct is then transformed to a probability to generate a confidence rating on a 0–1 scale:

$$\text{Confidence}=\frac{1}{1+\exp\left(-\mathrm{LO}_{\mathrm{correct}}^{\mathrm{total}}\right)}$$

Extensions of the Bayesian model. *Temporal weighting.* We considered the possibility that subjects may apply differential weights to pre- and post-decision motion when computing confidence[6,7]. To capture this possibility, we introduced free parameters w_{pre} and w_{post} that controlled the relative weights applied to pre- and post-decision evidence:

$$\mathrm{LO}_{\mathrm{correct}}^{\mathrm{total}}=w_{\mathrm{pre}}\mathrm{LO}_{\mathrm{correct}}^{\mathrm{pre}}+w_{\mathrm{post}}\mathrm{LO}_{\mathrm{correct}}^{\mathrm{post}}$$

Choice weighting. We considered that subjects might pay selective attention to post-decision evidence dependent on whether it is consistent/inconsistent with their initial choice (a form of commitment bias; this is similar to the "selective reduced-gain" model of Bronfman et al.[6]). To capture such effects, we introduced two weighting parameters w_{con} and w_{incon} that differentially weight confirmatory and disconfirmatory post-decision evidence:

If $\operatorname{sign}\left(\mathrm{LO}_{\mathrm{dir}}^{\mathrm{post}}\right)=\operatorname{sign}(a)$:

$$\mathrm{LO}_{\mathrm{correct}}^{\mathrm{total}}=\mathrm{LO}_{\mathrm{correct}}^{\mathrm{pre}}+w_{\mathrm{con}}\mathrm{LO}_{\mathrm{correct}}^{\mathrm{post}}$$

Otherwise:

$$\mathrm{LO}_{\mathrm{correct}}^{\mathrm{total}}=\mathrm{LO}_{\mathrm{correct}}^{\mathrm{pre}}+w_{\mathrm{incon}}\mathrm{LO}_{\mathrm{correct}}^{\mathrm{post}}$$

Choice bias. A second variant of commitment bias operates to boost confidence in the chosen response without altering sensitivity to post-decision evidence (the choice acts as a prior on subsequent confidence formation[25]; this is similar to Bronfman et al.'s "value-shift" model[6]). To capture such effects, we introduced a parameter b that modulated final confidence dependent on the choice:

$$\mathrm{LO}_{\mathrm{dir}}^{\mathrm{bias}}=\operatorname{sign}(a)\times\log\left(\frac{b}{1-b}\right)$$

$$\mathrm{LO}_{\mathrm{dir}}^{\mathrm{total}}=\mathrm{LO}_{\mathrm{dir}}^{\mathrm{pre}}+\mathrm{LO}_{\mathrm{dir}}^{\mathrm{post}}+\mathrm{LO}_{\mathrm{dir}}^{\mathrm{bias}}$$

If $a=1$:

$$\mathrm{LO}_{\mathrm{correct}}^{\mathrm{total}}=\mathrm{LO}_{\mathrm{dir}}^{\mathrm{total}}$$

Otherwise:

$$\mathrm{LO}_{\mathrm{correct}}^{\mathrm{total}}=-\mathrm{LO}_{\mathrm{dir}}^{\mathrm{total}}$$

Nonlinear confidence mapping. The ideal-observer model assumes that subjects faithfully report probability correct, which maximizes the quadratic scoring rule (QSR). We also considered the possibility that subjects may misperceive the scoring rule (or, equivalently, apply a nonlinear mapping between probability correct and reported confidence), with consequences for how particular confidence ratings were selected. For instance, subjects may overweight the extremes of the scale because they perceive these extremes as returning greater reward.

Such misperceptions can be captured by allowing a flexible mapping between the model's confidence and reported confidence. We implemented a one-parameter scaling of log-odds[22] that is able to capture both under- and overweighting of extreme confidence ratings:

$$\mathrm{LO}(\pi(c))=\gamma log\left(\frac{c}{1-c}\right)$$

$$\text{Confidence}=\frac{1}{1+\exp(-\mathrm{LO}(\pi(c)))}$$

where c denotes the interim output of the model's estimate of probability correct.

When $\gamma=1$, $\pi(c)=c$, and there is no distortion. When $\gamma>1$, the curve relating model confidence to reported confidence is S-shaped, whereas when $0<\gamma<1$, an inverted-S-shaped curve is obtained.

Informing confidence with decision time. Finally, we considered the possibility that in all models subjects may use decision time from the initial decision as a cue to confidence[23]. To capture this possibility, we modulated the final $\mathrm{LO}_{\mathrm{correct}}^{\mathrm{total}}$ of both the Bayesian and extended models by response time via a free parameter β_{RT}:

$$\text{LO}^{\text{total}}_{\text{correct}} \leftarrow \text{LO}^{\text{total}}_{\text{correct}} + \beta_{\text{RT}} log\,(\text{RT})$$

In the case of the mapping model the modulation by decision time was applied before passing $\text{LO}^{\text{total}}_{\text{correct}}$ through the nonlinear mapping function.

This set of model extensions led to a factorial combination of 5 model variants (ideal Bayesian, temporal weighting, choice weighting, choice bias, mapping) × 2 (non-response-time dependent, response-time dependent) = 10 models, which were fitted to each subject and dataset as described below.

Model fitting. We used Markov chain Monte Carlo methods implemented in STAN[55] to sample from posterior distributions of parameters given motion directions d, motion coherences θ_{pre} and θ_{post}, subjects' choices a and confidence ratings r.

Pseudo-code for the Bayesian model is given below (following the STAN convention, scale parameters are written as s.d.):

Priors:

$$m \sim N\,(0, 10)$$

$$k \sim N\,(0, 10)$$

Model:

$$X_{\text{pre}} \sim N\,(dk\theta_{\text{pre}}, 1)$$

$$X_{\text{post}} \sim N\,(dk\theta_{\text{post}}, 1)$$

$$a \sim \text{Bernoulli\_logit}\,(100(X_{\text{pre}} - m))$$

$$r \sim N\,(\text{conf}, 0.025)$$

Here "conf" is the output of the confidence computation detailed above. The logit function implements a steep softmax relating X_{pre} to a and is applied for computational stability. The mapping between model confidence and observed confidence allowed a small degree of imprecision ($\sigma = 0.025$) in subjects' ratings, roughly equivalent to grouping continuous ratings made on a 0–1 scale into ten bins.

We placed weakly informative priors over coefficients in the extended models for computational stability. In the weighted models, w parameters were drawn from $N\,(1, 1)$ distributions bounded below by 0 and above by 5. In the bias model, b was drawn from a uniform [0 1] distribution. In the nonlinear mapping model, γ was drawn from a positively constrained $N\,(1, 1)$ distribution. In the RT models, β_{RT} was drawn from an $N\,(0, 10)$ distribution.

We fitted each model with 12,000 samples divided across 3 chains separately for each subject's fMRI and behavioral datasets. We discarded 1,000 samples per chain for burn-in, resulting in 9,000 stored samples. Chains were visually checked for convergence and Gelman and Rubin's potential scale reduction factor $\hat{R}$ was calculated for all parameters[63]. For most models and subjects (469 out of 470), $\hat{R}$ values were all < 1.1, indicating good convergence. The fit of the choice-weighted + RT model to the behavioral session data failed to converge for one subject; this log-likelihood value was omitted from the model comparison calculations detailed below.

Model comparison. To compare models, we assessed the ability of a model fit to behavioral data to capture the data of the same subject in the fMRI session, and vice-versa. For each subject and model, we drew 1,000 samples from posterior distributions of fitted parameters and generated synthetic choice and confidence data. The trialwise log-likelihood (itself a sum of choice and confidence rating log-likelihoods) was summed across trials and stored for each parameter draw, and then averaged across draws to return a subject- and model-specific cross-validated log-likelihood. Fitted parameter values from the best-fitting Bayesian + RT model for behavioral and fMRI sessions are listed in Supplementary Table 3.

Model simulations. To visualize qualitative features of the Bayesian model (Fig. 1b), we simulated 10,000 trials from each condition of the factorial design with $k = 4$ and $m = 0$. Pre- and post-decision motion coherences were crossed in a fully factorial design and drawn from the set 0%, 25% or 50%. True motion direction d was selected randomly on each trial.

To determine the ability of the best-fitting Bayesian + RT model to account for subjects' choices and confidence ratings (a posterior predictive check), we drew 1,000 samples from posterior distributions of fitted parameters and for each draw simulated one trial sequence with these parameter settings and averaged over simulations. To obtain regressors for fMRI and mediation analyses, we also stored values of pre-decision evidence ($\text{LO}^{\text{pre}}_{\text{correct}}$) and post-decision evidence ($\text{LO}^{\text{post}}_{\text{correct}}$) averaged over 5,000 trials per condition (3 pre-decision coherence levels × 3 post-decision coherence levels × 2 choice accuracies).

fMRI acquisition and preprocessing. Whole-brain fMRI images were acquired using a 3 T Allegra scanner (Siemens) with an NM011 head transmit coil (Nova Medical, Wakefield, MA) at New York University's Center for Brain Imaging. BOLD-sensitive echo-planar images (EPI) were acquired using a Siemens epi2d BOLD sequence (42 transverse slices, TR = 2.34 s; echo time = 30 ms; 3 × 3 × 3 mm resolution voxels; flip angle = 90 degrees; 64 × 64 matrix; slice tilt −30 deg $T > C$; interleaved acquisition). The main experiment consisted of 4 runs of 315 volumes, and the localizer scan consisted of a single run of 211 volumes. A high-resolution T1-weighted anatomical scan (MPRAGE, 1 × 1 × 1 mm voxels, 176 slices) and local field maps were also acquired.

All preprocessing was carried out using SPM12 v6225 (Statistical Parametric Mapping; http://www.fil.ion.ucl.ac.uk/spm). The first five volumes of each run were discarded to allow for T1 equilibration. Functional images were slice-time corrected, realigned and unwarped using the collected field maps[56]. Structural T1-weighted images were coregistered to the mean functional image of each subject using the iterative mutual information-based algorithm. Each participant's structural image was segmented into gray matter, white matter and cerebral spinal fluid images using a nonlinear deformation field to map it onto a template tissue probability map[57]. These deformations were applied to both structural and functional images to create new images spatially normalized to MNI space and interpolated to 2 × 2 × 2 mm voxels. Normalized images were spatially smoothed using a Gaussian kernel with a full-width half-maximum of 6 mm.

fMRI analysis. We employed a combination of region-of-interest (ROI) analyses on trial-by-trial activity estimates, multilevel mediation models and standard whole-brain general linear model (GLM) approaches.

Whole-brain univariate analysis. We used SPM12 for first-level analyses. In all GLMs, regressors were convolved with a canonical hemodynamic response function. Motion correction parameters estimated from the realignment procedure and their first temporal derivatives were entered as nuisance covariates, and low-frequency drifts were removed using a high-pass filter (128-s cutoff).

GLM1. GLM1 was constructed to examine activity associated with changes in post-decision motion strength. Correct and incorrect trials were modeled as separate stick functions time-locked to the onset of the post-decision motion plus parametric modulations by post-decision motion strength (low = −1, medium = 0, high = 1). Additional regressors were also included at the onset of pre-decision motion (parametrically modulated by pre-decision motion strength and log response times) and confidence rating period.

GLM2. GLM2 was constructed to examine activity associated with changes in reported confidence. A stick function time-locked to confidence rating onset was parametrically modulated by reported confidence. Regressors were also included at the onset of pre-decision motion (parametrically modulated by log response times) and post-decision motion.

ROI analysis. A priori regions of interest were specified as follows. The pMFC ROI was an 8-mm sphere around peak coordinates (MNI coordinates [x y z] = [0 17 46]) obtained from our previous study of decision confidence[12]. Anterior prefrontal ROIs were obtained from the right-hemisphere atlas developed by Neubert et al.[24] (area 46, FPl and FPm) and mirrored to the left hemisphere to create bilateral masks. The vmPFC ROI was an 8-mm sphere around peak coordinates [−1 46 −7] obtained from a meta-analysis of value-related activity[58]. The ventral striatum ROI was specified anatomically from the Oxford-Imanova Striatal Structural Atlas included with FSL (http://fsl.fmrib.ox.ac.uk). Within each ROI we averaged single-trial β estimates over voxels, scaled the time series to have zero mean and unit s.d., and computed the mean activity per condition.

Quantification of single-trial response magnitudes. To facilitate both ROI and mediation analyses, we estimated single-trial BOLD responses as a β time series. This was achieved by specifying a GLM design matrix with separate regressors (stick functions) for each trial, each aligned to either the onset of the post-decision motion stimulus (for PDE analyses in Fig. 3) or the confidence rating period (for mediation models and regressions on confidence; Figs. 4 and 5). Each regressor was convolved with a canonical hemodynamic response function. Motion-correction parameters estimated from the realignment procedure and their first temporal derivatives were entered as nuisance covariates, and low-frequency drifts were removed using a high-pass filter (128 s cutoff). One important consideration in using single-trial estimates is that the β for a given trial can be strongly affected by acquisition artifacts that cooccur with that trial (for example, motion or scanner pulse artifacts). For each subject we therefore computed the grand mean β estimate across both voxels and trials, and excluded any trial whose mean β estimate across voxels exceeded 3 s.d. from this grand mean[38]. An average of 3.6 trials per subject (1.0%; maximum = 9 trials) were excluded.

To visualize the relationship between activity and task variables over time, we also extracted the preprocessed BOLD data per TR. Low-frequency drifts (estimated using a cosine basis set, 128 s cutoff) and motion parameters plus their first temporal derivatives were regressed out of the signal, and the residual activity was oversampled at 10 Hz. Time courses were extracted from 12-s windows time-locked to the onset of pre-decision motion. To construct Fig. 4c, we applied a GLM (see below) to each time point, resulting in a time course of β weights for each regressor. Nonparametric permutation tests were used to assess group-level significance of β weights. For each permutation, we randomized the assignment between BOLD time series and trial labels and recalculated the group-level t-statistic comparing β weights against zero (10,000 permutations). Individual time points were labeled as significant if the true t-statistic fell outside the 2.5th or 97.5th percentiles of the null distribution.

ROI GLMs. As in our regression analyses of behavior, we modeled subject-level slopes and intercepts, and report coefficients and statistics at the population level. To test for an interaction between response accuracy and post-decision evidence, we fitted the following model to each ROI β series:

$$\begin{aligned}&\mathrm{BOLD} \sim \mathrm{accuracy} + \mathrm{pre\_decision\_coherence} + \mathrm{post\_decision\_}\\&\mathrm{coherence} + \mathrm{accuracy} \times \mathrm{pre\_decision\_coherence} + \mathrm{accuracy} \times\\&\mathrm{post\_decision\_coherence} + \log(\mathrm{RT})\end{aligned}$$

Accuracy was specified as error = −1, correct = 1; pre- and post-decision coherence were specified as low = −1, medium = 0, high = 1.

To estimate relationships between ROI activity and pre- and post-decision evidence from the fitted computational model (i.e., log-odds correct) we fitted the following model:

$$\mathrm{BOLD} \sim \mathrm{LO}^{\mathrm{pre}}_{\mathrm{correct}} + \mathrm{LO}^{\mathrm{post}}_{\mathrm{correct}} + \log(\mathrm{RT})$$

To assess relationships between confidence and activity on both change-of-mind and no-change-of-mind trials, we conducted a segmented regression analysis. This method partitions the independent variable into discrete intervals, and a separate slope is fit to each interval. Here, we separated the effect of confidence on change (confidence ≤ 0.5) and no-change (confidence > 0.5) trials, and fit the following model:

$$\mathrm{BOLD} \sim \mathrm{change\_confidence} + \mathrm{no\_change\_confidence} + \mathit{log}\,(\mathrm{RT})$$

Multilevel mediation analysis. We performed multilevel mediation analysis of a standard three-variable model[20] using the Mediation Toolbox (http://wagerlab.colorado.edu/tools). Mediation analysis assesses whether covariance between two variables (X and Y) is explained by a third variable (the mediator, M). Significant mediation is obtained when inclusion of M in a path model of the effects of X on Y significantly alters the slope of the X–Y relationship. When applied to fMRI data, mediation analysis thus extends the standard univariate model by incorporating an additional outcome variable (in this case, confidence reports) and jointly testing three effects of interest: the impact of X (post-decision evidence, $\mathrm{LO}^{\mathrm{post}}_{\mathrm{correct}}$) on brain activity (path a); the impact of brain activity on Y (confidence reports), controlling for X (path b); and formal mediation of X on Y by brain activity M. In all models we included log reaction times and pre-decision evidence ($\mathrm{LO}^{\mathrm{pre}}_{\mathrm{correct}}$) as covariates of no interest.

The Mediation Toolbox permits a multilevel implementation of the standard mediation model, treating participant as a random effect[59]. Significance estimates for paths a, b and $a \times b$ are computed through bootstrapping. We estimated distributions of subject-level path coefficients by drawing 10,000 random samples with replacement. Two-tailed P-values were calculated at each voxel/ROI from the bootstrap confidence interval[60].

Whole-brain statistical inference. Single-subject contrast images were entered into a second-level random effects analysis using one-sample t-tests against zero to assess group-level significance. To correct for multiple comparisons, we used Gaussian random field theory as implemented in SPM12 to obtain clusters satisfying $P < 0.05$, family-wise error (FWE)-corrected at a cluster-defining threshold of $P < 0.001$. Numerical simulations and tests of empirical data collected under the null hypothesis show that this combination of cluster-defining threshold and random field theory produces appropriate control of false positives[61,62].

To apply multiple-comparisons correction to the multilevel mediation model output, we took a non-parametric approach because second-level images already comprise bootstrapped P-values. The cluster extent threshold for FWE correction was estimated on the basis of Monte Carlo simulation (100,000 iterations) using the 3dClustSim routine in AFNI (version compiled September 2015; http://afni.nimh.nih.gov) and SPM12's estimate of the intrinsic smoothness of the residuals. Again, this method in conjunction with a cluster-defining threshold of $P < 0.001$ provides appropriate control over false positives[61,62]. Statistical maps were visualized using FSLview (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki) and Surf Ice (https://www.nitrc.org/projects/surfice/).

Life Sciences Reporting Summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

Code availability. MATLAB, STAN and R code for reproducing all analyses and computational model fits is available on GitHub (https://github.com/metacoglab/FlemingVdPuttenDaw).

Data availability. Anonymized behavioral data are available on GitHub (https://github.com/metacoglab/FlemingVdPuttenDaw). Unthresholded group-level statistical maps are available on NeuroVault (https://neurovault.org/collections/VEJNEJRA/). Other data that support the findings of this study are available from the corresponding author upon reasonable request.

References

47. Brainard, D. H. The Psychophysics Toolbox. *Spat. Vis.* **10**, 433–436 (1997).
48. Pelli, D. G. The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spat. Vis.* **10**, 437–442 (1997).
49. Roitman, J. D. & Shadlen, M. N. Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *J. Neurosci.* **22**, 9475–9489 (2002).
50. Staël von Holstein, C.-A. S. Measurement of subjective probability. *Acta Psychol. (Amst.)* **34**, 146–159 (1970).
51. Schotter, A. & Trevino, I. Belief elicitation in the laboratory. *Annu. Rev. Econom* **6**, 103–128 (2014).
52. Moore, D. A. & Healy, P. J. The trouble with overconfidence. *Psychol. Rev.* **115**, 502–517 (2008).
53. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. Preprint at https://arxiv.org/abs/1406.5823 (2014).
54. Fox, J. & Weisberg, S. *An R Companion to Applied Regression* (Sage, Thousand Oaks, CA, USA, 2011).
55. Carpenter, B. et al. Stan: a probabilistic programming language. *J. Stat. Softw.* **20**, 1–37 (2016).
56. Andersson, J. L., Hutton, C., Ashburner, J., Turner, R. & Friston, K. Modeling geometric deformations in EPI time series. *Neuroimage* **13**, 903–919 (2001).
57. Ashburner, J. & Friston, K. J. Unified segmentation. *Neuroimage* **26**, 839–851 (2005).
58. Bartra, O., McGuire, J. T. & Kable, J. W. The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Neuroimage* **76**, 412–427 (2013).
59. Wager, T. D., Davidson, M. L., Hughes, B. L., Lindquist, M. A. & Ochsner, K. N. Prefrontal-subcortical pathways mediating successful emotion regulation. *Neuron* **59**, 1037–1050 (2008).
60. Efron, B. & Tibshirani, R. *An Introduction to the Bootstrap* (CRC Press, Boca Raton, FL, USA, 1993).
61. Woo, C.-W., Krishnan, A. & Wager, T. D. Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. *Neuroimage* **91**, 412–419 (2014).
62. Eklund, A., Nichols, T. E. & Knutsson, H. Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. *Proc. Natl. Acad. Sci. USA* **113**, 7900–7905 (2016).
63. Gelman, A. & Rubin, D. B. Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7**, 457–472 (1992).

nature research

Corresponding author(s): Stephen Fleming

□ Initial submission □ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

No statistical tests were used to predetermine the sample size, but this sample size is within the standard range in the field.

2. Data exclusions

Describe any data exclusions.

Established exclusion criteria (Online Methods, section "Participants") were applied for excessive head motion (n=1) and lack of variability in confidence (n=1).

3. Replication

Describe whether the experimental findings were reliably reproduced.

Computational and behavioural findings were reliably reproduced across behavioural and fMRI testing sessions. All attempts at replication were successful.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

Participants were not grouped and hence no randomization was performed. Trial order was fully randomized for each subject.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Data collection and analysis were not performed blind to the conditions of the experiments

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
□	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	A statement indicating how many times each experiment was replicated
□	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
□	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
□	☒	The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
□	☒	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
□	☒	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

The task was programmed in MATLAB 2014b using Psychtoolbox (version 3.0.12).

Behavioural data and fMRI ROI data were analysed using hierarchical mixed-effects regression using lme4 in R (Version 3.3.3). P-values for linear regression coefficients were obtained using the car package in R (version 2.1) as Wald type III chi-squared tests. Computational models were implemented in STAN (rstan, Version 2.6.0).

fMRI data were analysed using SPM 12 (v6225), AFNI (version compiled September 2015) and custom scripts in MATLAB. Mediation analyses were carried out with the Mediation Toolbox in MATLAB (https://canlabweb.colorado.edu/wiki/doku.php/help/mediation/m3_mediation_fmri_toolbox). MRI images were visualised using FSL (version 5.0.8) and SurfIce.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

All unique materials are readily available from the authors or freely available online.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

No antibodies were used

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

No cell lines were used

b. Describe the method of cell line authentication used.

No cell lines were used

c. Report whether the cell lines were tested for mycoplasma contamination.

No cell lines were used

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

No cell lines were used

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

No animals were used

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

Twenty-five healthy participants were included in the analysis of behavioural data (14 females, mean age 24.0, SD = 3.6); of these, twenty-two healthy participants were included in the analysis of fMRI data (12 females, mean age 24.1, SD = 3.4).

nature research

Corresponding author(s): Stephen Fleming

☐ Initial submission ☐ Revised version ☒ Final submission

MRI Studies Reporting Summary

Form fields will expand as needed. Please do not leave fields blank.

Experimental design

1. Describe the experimental design.

Event-related, randomized trial sequence

2. Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.

4 blocks, 90 trials per block, 5.9s per trial, 2s inter-trial interval

3. Describe how behavioral performance was measured.

Button press, response time, confidence rating. Performance was assessed via hierarchical mixed-effects regression of the effects of motion coherence on accuracy and confidence.

Acquisition

4. Imaging

a. Specify the type(s) of imaging.

Functional and structural

b. Specify the field strength (in Tesla).

3 Tesla

c. Provide the essential sequence imaging parameters.

BOLD-sensitive echo-planar images (EPI) were acquired using a Siemens epi2d BOLD sequence (42 transverse slices, TR = 2.34s; echo time = 30ms; 3 x 3 x 3 mm resolution voxels; flip angle = 90 degrees; 64 x 64 matrix; slice tilt -30deg T > C; interleaved acquisition).

d. For diffusion MRI, provide full details of imaging parameters.

N/A

5. State area of acquisition.

Whole-brain

Preprocessing

6. Describe the software used for preprocessing.

SPM12 v6225

7. Normalization

a. If data were normalized/standardized, describe the approach(es).

Each participant's structural image was segmented into gray matter, white matter and cerebral spinal fluid images using a nonlinear deformation field to map it onto a template tissue probability map. These deformations were applied to both structural and functional images to create new images spatially normalized to Montreal Neurological Institute space and interpolated to 2x2x2 mm voxels.

b. Describe the template used for normalization/transformation.

SPM12's MNI template

8. Describe your procedure for artifact and structured noise removal.

Motion correction parameters estimated from the realignment procedure and their first temporal derivatives were entered as nuisance covariates.

9. Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.

N/A

Statistical modeling & inference

10. Define your model type and settings.

First-level mass univariate; second-level random effects

11. Specify the precise effect tested.

Positive/negative interaction of post-decision motion strength x choice accuracy (GLM1); positive/negative parametric effects of confidence (GLM2).

12. Analysis

a. Specify whether analysis is whole brain or ROI-based.

Whole-brain and ROI-based

b. If ROI-based, describe how anatomical locations were determined.

The pMFC ROI was an 8mm sphere around peak coordinates (MNI coordinates [x, y, z] = [0 17 46]) obtained from Fleming et al. (2012). Anterior prefrontal ROIs were obtained from the right-hemisphere atlas of Neubert et al. (2014) and mirrored to the left hemisphere to create bilateral masks (area 46, FPm, FPl). The vmPFC ROI was an 8mm sphere around peak coordinates [-1 46 -7] obtained from a meta-analysis of value-related activity (Bartra et al. 2013). The ventral striatum ROI was specified anatomically from the Oxford-Imanova Striatal Structural atlas included with FSL.

13. State the statistic type for inference. (See Eklund et al. 2016.)

For all analyses except multilevel mediation, statistical inference was conducted using Gaussian random field theory as implemented in SPM12 to obtain clusters satisfying P<0.05, family-wise error (FWE) corrected at a cluster-defining threshold of P<0.001 uncorrected. To apply multiple comparisons correction to the multilevel mediation model output we took a non-parametric approach due to second-level images already comprising bootstrapped P-values. The cluster extent threshold for FWE correction was estimated based on Monte Carlo simulation (100,000 iterations) using the 3dClustSim routine in AFNI (version compiled September 2015), cluster-defining threshold P<0.001 uncorrected. Numerical simulations and tests of empirical data collected under the null hypothesis show that both methods provide appropriate control over false positives (Eklund et al. 2016).

14. Describe the type of correction and how it is obtained for multiple comparisons.

FWE and Monte-Carlo

15. Connectivity

a. For functional and/or effective connectivity, report the measures of dependence used and the model details.

N/A

b. For graph analysis, report the dependent variable and functional connectivity measure.

N/A

16. For multivariate modeling and predictive analysis, specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.

N/A