Domain-Specific Patterns of Activity Support Metacognition in Human Prefrontal Cortex

Jorge Morales[1*], Hakwan Lau[2,3], Stephen M. Fleming[4]

[1] Department of Philosophy, Columbia University, New York, NY, USA. jorge.morales@columbia.edu
[2] Department of Psychology, UCLA, Los Angeles, CA, USA. hlau@psych.ucla.edu
[3] Brain Research Institute, UCLA
[4] Wellcome Trust Centre for Neuroimaging, University College London, London, United Kingdom. stephen.fleming@ucl.ac.uk

* *Corresponding author:* jorge.morales@columbia.edu (J. Morales)

Abstract: Metacognition is the capacity to evaluate the success of one's own cognitive processes in various domains, e.g. memory and perception. It remains controversial whether metacognition relies on a domain-general resource that is applied to different tasks, or whether self-evaluative processes are domain-specific. Here we directly investigated this issue by examining the neural substrates engaged when metacognitive judgments were made during perceptual and memory tasks matched for stimulus and performance characteristics. By comparing patterns of functional magnetic resonance imaging (fMRI) activity while subjects evaluated their performance, we revealed both domain-specific and domain-general metacognitive representations. Multi-voxel activity patterns in right frontopolar cortex predicted levels of confidence in a domain-specific fashion, whereas domain-general signals predicting confidence and accuracy were found in a widespread network in the frontal and posterior midline. Moreover, individual differences in perceptual and memory metacognitive ability were predicted by the degree of confidence decoding in frontal and parietal cortex, respectively. The demonstration of domain-specific metacognitive representations suggests the presence of a content-rich mechanism available to introspection and cognitive control.

Keywords: metacognition; memory; perception; multi-voxel pattern analysis; fMRI; prefrontal cortex

Major subject area: Neuroscience

INTRODUCTION

Metacognition is the capacity to evaluate the success of one's cognitive processes in various domains, e.g. perception or memory (1-4). Metacognitive ability can be assessed in the laboratory by quantifying the trial-by-trial correspondence between objective performance and subjective confidence (5-8). Anatomical (9-11), functional (12-16) and neuropsychological (1-4,17-19) evidence indicates specific neural substrates (especially in frontolateral, frontomedial, and parietal regions) contribute to metacognition across a range of task domains, including memory and perception. However, the neurocognitive architecture supporting metacognition remains controversial. Does metacognition rely on a common, domain-general resource that is recruited to evaluate performance on a variety of tasks? Or is metacognition supported by domain-specific components?

Current computational perspectives (20,21) suggest both domain-general and domain-specific representations may be important for guiding behavior. On the one hand, one needs to be able to compare confidence estimates in a "common currency" across a range of arbitrary decision scenarios (22). One solution to this problem is to maintain a global resource with access to arbitrary sensorimotor mappings (23-25). Candidate neural substrates for a domain-general resource are the frontoparietal and cingulo-opercular networks, known to be involved in arbitrary control operations (23). In particular, the dorsomedial prefrontal cortex (dmPFC; encompassing the paracingulate cortex and pre-supplementary motor area) has been implicated both in representing confidence and detecting errors across a range of tasks (12,26). On the other hand, if the system only had access to generic confidence signals, appropriate switching between particular tasks or strategies on the basis of their expected success would be compromised. Functional imaging evidence implicates human frontopolar cortex in tracking the reliability of specific alternative strategies during decision-making (27), and such regions may also support domain-specific representations of confidence.

Current behavioral evidence of a shared resource for metacognition is ambiguous, in part due to the difficulty of distilling metacognitive processes from those supporting primary task performance (5-7). Some studies have found that efficient metacognition in one task predicts good metacognition in another (9,28,29), whereas others indicate the independence of metacognitive abilities (14,30,31). Recent studies employing bias-free measures of metacognition have identified differences in the neural correlates of memory and perceptual metacognition in both healthy subjects (9,14) and neuropsychological patients (19). However, the study of behavioral individual differences provides only indirect evidence of the neural and computational architecture supporting metacognition.

Here, we directly investigate this ontology by examining neural substrates engaged when metacognitive judgments are made during perceptual and memory tasks matched for stimulus and performance characteristics. We employ a combination of univariate and multivariate analyses of functional magnetic resonance imaging (fMRI) data to identify domain-specific and domain-general neural substrates engaged during metacognitive judgments. We also distinguish

activations engaged by a metacognitive judgment from neural activity which parametrically tracks confidence level. Together, our findings reveal generic and specific confidence representations co-exist, consistent with a computational hierarchy underpinning effective metacognition.

RESULTS

We analyzed the data of 24 subjects who underwent hemodynamic neuroimaging while performing two-alternative forced-choice discrimination tasks in perceptual (PER) and memory (MEM) domains (Figure 1A). In the perceptual task, subjects were asked to indicate the brighter of two stimuli (words or abstract shapes). In the memory task, subjects were asked to memorize exemplars of the same stimulus types, and then select the previously-learned stimulus from two stimuli presented on each trial. By using a 2×2 design (TASK DOMAIN × STIMULUS TYPE), we were able to ensure differences in neural activations across perception and memory tasks were not confounded by stimulus type. In half of the trials (Confidence condition) subjects performed a metacognitive evaluation after the discrimination task by rating their confidence in the correctness of their response by selecting a number from a 1-to-4 scale (1=not confident; 4=very confident). In order to differentiate metacognitive judgment-related activity from visuomotor activity engaged by use of the confidence scale, in the other half of the trials (Follow condition) subjects were asked to respond according to a highlighted number without evaluating confidence in their response.

Behavior

We first compared task performance, measured by percentage of correct responses, across condition, task, and stimulus type. A 2×2×2 repeated measures ANOVA (CONFIDENCE/FOLLOW × PERCEPTION/MEMORY × SHAPES/WORDS) showed that performance was well-matched across rating conditions (Confidence vs Follow) ($F_{1,23}$=3.036, P=0.095; Figure S1). Matching performance across stimulus type was more difficult to achieve because subjects' memory for words was expected to be considerably higher than that for abstract shapes trials based on pilot data (see Methods for details). Instead, we aimed to match subjects' performance for each stimulus type across task domains by titrating the difficulty of the perceptual task to approximate the performance expected for each stimulus type in the memory task (shapes: PER M=73%, MEM M=67%; words: PER M=81%, MEM M=89%; Figure 1B). Critically, this ensured performance was matched across task domains when averaging stimulus types across participants (PER: M=77%, MEM: M=78%; paired t-test T_{23}=0.38, P=0.70; Figure 1C).

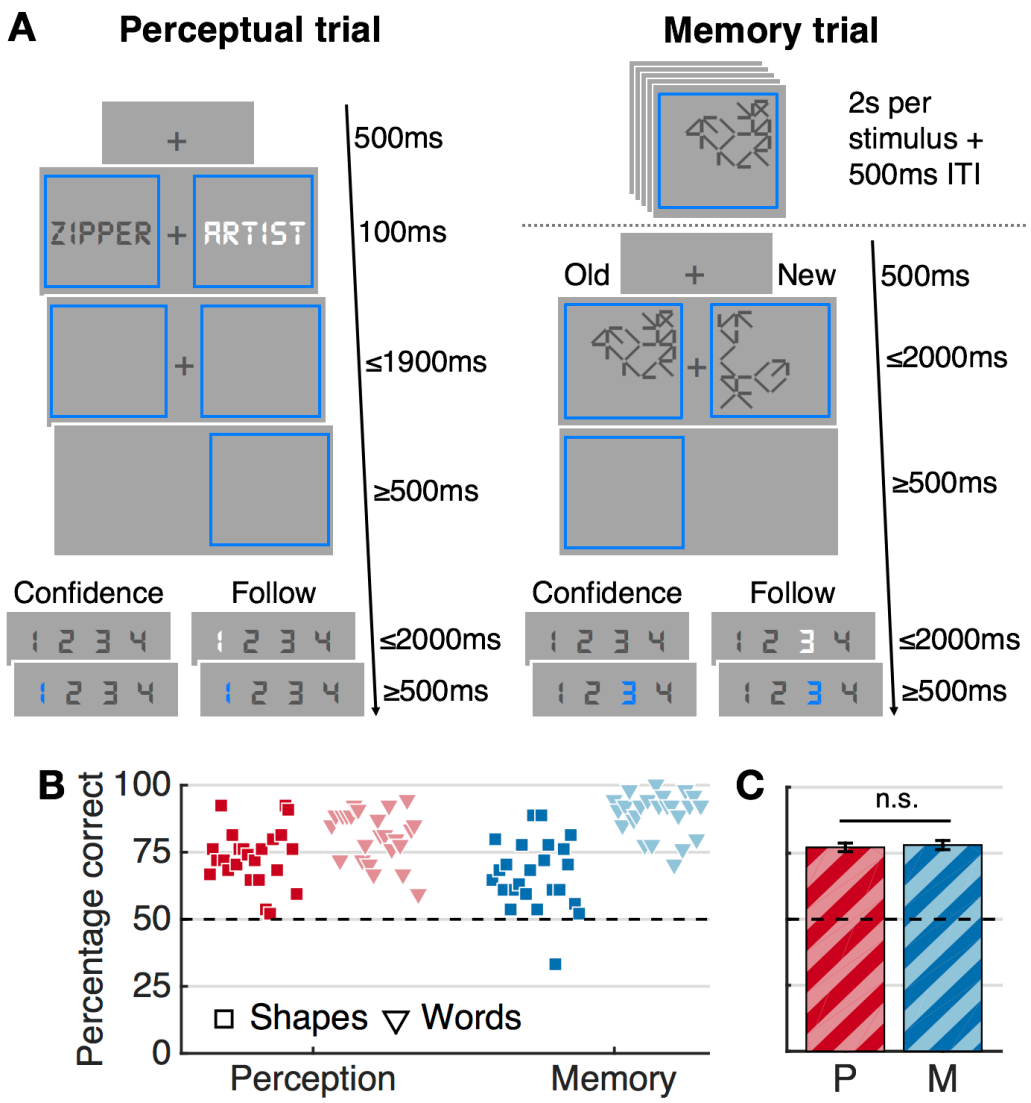

Figure 1. Task design & performance results. (A) Subjects performed two-alternative forced-choice discrimination tasks about perception and memory. In perception blocks, subjects selected the brighter of two stimuli. Memory blocks started with an encoding period and then subjects indicated in each trial which of two stimuli appeared during the encoding period. Abstract shapes and words were used as stimuli in both tasks. In Confidence blocks, subjects rated their confidence and in Follow blocks they pressed the highlighted number. **(B)** Scatter plot of percentage correct responses per block type in the Confidence condition. Each marker represents a subject. **(C)** Mean percentage correct responses by domain, averaged over subjects and stimulus types. Dotted lines indicate chance performance. Bars indicate standard error of the mean (s.e.m.). See Figure S1 for performance in the Follow condition. P=perception M=memory

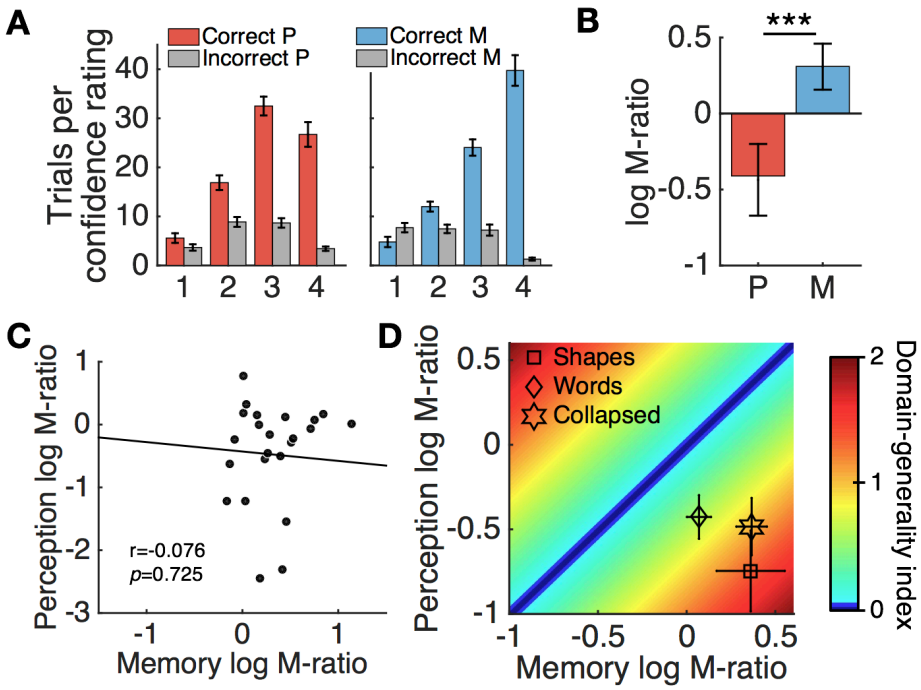

Figure 2. Metacognitive measures. (A) Mean number of correct and incorrect trials per confidence rating. **(B)** Metacognitive efficiency measured by log M-ratio (meta-*d*'/*d*'). Zero indicates that metacognitive sensitivity (meta-*d*') is equal to task sensitivity *d*' (i.e. the *d*' that would have been predicted to give rise to the observed confidence rating data assuming a signal detection theoretic ideal observer). Group-level hierarchical Bayesian estimates differed significantly between domains. Error bars indicate 95% high-density interval (HDI) from posterior samples. **(C)** Subjects' mean individual hierarchical Bayesian estimates of metacognitive efficiency scores were not correlated across the perceptual and memory domains. **(D)** To estimate the degree of generality of metacognitive efficiency between domains, we calculated a domain-general index (DGI) for each subject that quantifies the similarity between their log M-ratio scores in each domain (see main text for details). Greater DGI scores indicate less metacognitive consistency across domains. Mean log M-ratios are shown for reference. Bars in (A) and (D) indicate s.e.m. *** $P_{\theta>0} \sim 1$. P=perception M=memory.

As expected, subjects gave higher confidence ratings after correct decisions than after incorrect decisions (Figure 2A), and mean confidence ratings were similar across task domains (PER M=2.62, MEM M=2.47; paired *t*-test T_{23}=1.26, *P*=0.22). Reaction times for confidence ratings were not different between domains (PER M=518ms, MEM M=516ms; paired *t*-test T_{23}=0.16, *P*=0.87). We next estimated log M-ratio, a metacognitive efficiency measure derived from signal detection theory that assays the degree to which confidence ratings distinguish between correct and incorrect trials (5,6,32). Metacognitive efficiency in the perceptual task was significantly lower than in the memory task ($P_{\theta>0} \sim 1$; see Figure 2B & Methods for details), consistent with

previous findings (19). We did not find a correlation between metacognitive efficiency scores in the perceptual and memory domains (r_{22}=-0.076; P=0.72; Figure 2C). To further assess a potential covariation between metacognitive abilities in each domain, we calculated for each subject a domain-generality index (DGI) that quantifies the similarity between their scores in each domain (19):

$$\mathrm{DGI} = |\log M_P - \log M_M|$$

where M_P = perceptual M-ratio and M_M = memory M-ratio. Lower DGI scores indicate more similar metacognitive efficiencies between domains (DGI=0 indicates identical scores). Mean DGI for shapes (1.42), words (0.66) and collapsed by stimulus type (0.95) were higher than zero (Figure 2D). Metacognition for words was behaviorally more stable across domains as the DGI was smaller than for shapes (paired *t*-test: T_{23}=2.86; P=0.009). Together, these results suggest domain-specific constraints on metacognitive ability.

fMRI analyses

We next turned to our fMRI data to assess the overlap between neural substrates engaged when metacognitive judgments are made during perceptual and memory tasks. In standard univariate analyses, we found elevated activity in dACC/pre-SMA and bilateral insulae when comparing the Confidence against the Follow condition (Figure S2A). We also replicated negative correlations between confidence and activation in dACC/pre-SMA, parietal cortex and PFC that have been reported in several previous studies (12-14,33) (Figure S2B). However, we reasoned that while metacognition-related activity may show a consistent univariate profile across tasks (see Figures S3-S5 for univariate analyses), distributed activity patterns in frontal and parietal areas may carry task-specific information (33,34). To investigate the domain-specificity of activity patterns in frontoparietal and cingulo-opercular regions we therefore performed a series of multivoxel pattern analyses (MVPA). We focused on two features of metacognition-related activity in the analysis of functional imaging data - judgment-related (JR) activity (the difference between Confidence trials requiring a metacognitive judgment and the visuomotor control condition) and confidence level-related (CLR) activity (the parametric relationship between confidence and activity). To separately quantify these effects for each task domain, we focused on four a priori regions of interest (ROI) in dACC/pre-SMA, bilateral rostrolateral prefrontal cortex (rlPFC) and precuneus (PCUN) (Figure S4), which previous studies have found to be recruited by perceptual and memory metacognition (9,10,12-14), together with whole-brain searchlight analyses.

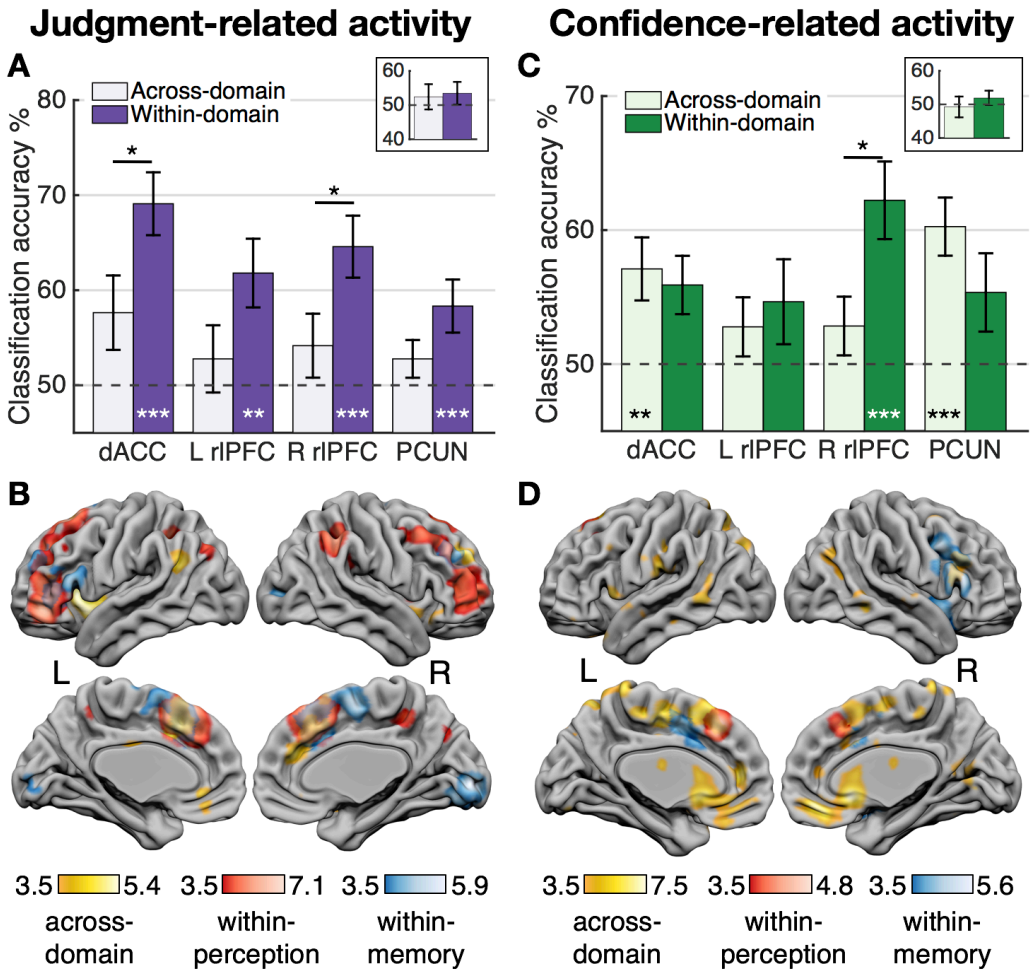

Figure 3. MVPA Results. *Judgment Related-Activity* **(A)** ROI results for across-domain (light colors) and mean within-domain (dark colors) classification accuracy of Confidence vs Follow trials. Inset: As a control, classification accuracy in the ventricles was not different from chance (across: T_{23}=0.66, P=0.52; within: T_{23}=1.04, P=0.31). **(B)** Searchlight analysis for same classifications as in (A) (within-domain classifications uncollapsed). *Confidence Level Related-Activity* **(C)** Same as in (A) but for low versus high confidence classification accuracy. Inset: Classification accuracy in the ventricles was not different from chance (across: T_{21}=-0.24, P=0.81; within: T_{21}=0.86, P=0.40). **(D)** Searchlight analysis for same classifications as in (C)**,** with within-domain classifications uncollapsed and visuomotor activity exclusively masked (see Figure S8 for visuomotor mask and Figure S9 for whole-brain maps of raw within-domain classification accuracy values). Bars in (A) and (C) indicate means and error bars s.e.m. Dashed lines indicate chance classification (50%). See Figures S6-S7 for ROI results split by domain. All clusters in (B) and (D) are significant at cluster-defining threshold P<0.001, corrected for multiple comparisons at P_{FWE}<0.05. Image displayed at P<0.001, uncorrected. Color bars indicate T-scores. See Tables S4-S6. L=left; R=right. *** P≤0.001 ** P≤0.01 * P<0.05.

ROI analysis of judgment-related activity. If metacognitive judgments are based on domain-general processes (i.e. shared across perceptual and memory tasks), a decoder trained to classify JR-activity in perceptual trials should accurately discriminate JR-activity when tested on memory trials (and vice versa). Alternatively, domain-specific activity profiles would be indicated by significant within-domain classification of JR-activity in the absence of across-domain transfer. To adjudicate between these hypotheses, we performed a support vector machine (SVM) decoding analysis using as input vectors the run-wise beta images pertaining to Confidence and Follow trials obtained from a general linear model (GLM). For within-domain classification we used standard leave-one-out cross-validations for each domain and we tested for across-domain generalization using a cross-classification (see Methods for details). Chance classification in both analyses was 50%.

Mean within-domain classification results were significantly above chance in all ROIs (one-sample *t*-tests Bonferroni corrected for multiple comparisons α=0.05/4=0.0125: dACC/pre-SMA T_{23}=5.77, P=6.99×10^{-6}; L rlPFC T_{23}=3.27, P=0.003; R rlPFC T_{23}=4.47, P=0.0002; PCUN T_{23}=2.98, P=0.007; dark colors in Figure 3A; see Figure S6 for results split by domain). In contrast, across-domain generalizations were not significantly different from chance in any ROI (one sample *t*-test Bonferroni corrected: dACC/pre-SMA, T_{23}=1.95, P=0.063; L rlPFC, T_{23}=0.76, P=0.44; R rlPFC, T_{23}=1.24, P=0.228; PCUN T_{23}=1.40, P=0.175; light colors in Figure 3A). This suggests that the patterns of activity that distinguish metacognitive judgments from the visuomotor control condition in one domain are distinct from similar patterns in the other domain. In particular, within-domain classification accuracy was significantly different from across-domain classification accuracy (dACC/pre-SMA: T_{23}=2.88, P=0.008: right rlPFC: T_{23}=2.24, P=0.035). These results are consistent with the hypothesis that metacognitive judgments recruit domain-specific patterns of cortical activity in anterior frontal and medial-parietal regions.

Searchlight analysis of judgment-related activity. We ran a similar decoding analysis using an exploratory whole-brain searchlight, obtaining a classification accuracy value per voxel (35). Consistent with our ROI results, we observed significant within-domain classification in large swathes of bilateral PFC for both perception (red) and memory (blue) (Figure 3B & Table S4). Within-perception classification was also successful in parietal regions—precuneus in particular—and within-memory activity patterns were classified accurately in occipital regions. We also identified clusters showing significant across-domain generalization (yellow) in dACC, pre-SMA, SFG (BA 9), supramarginal gyrus (BA 40), and bilateral IFG/insula, consistent with univariate results (Figure S2A).

ROI analysis of confidence level-related activity. We next asked whether confidence is encoded in a domain-general or domain-specific fashion by applying a similar approach to discriminate low versus high confidence trials. Note that in this case, and unlike for judgment-related activity (Figure S5B), ROI univariate analyses did not reveal any differences in confidence-related activity between domains (Figure S5A). We hypothesized that if confidence level is encoded by

domain-general neural activity patterns, it should be possible to train a decoder to discriminate low (1-2) from high (3-4) confidence rating patterns in the perceptual task and then accurately classify confidence on the memory task (and vice versa). In the absence of across-domain classification, within-domain classification above chance is indicative of confidence level-related domain-specific activity patterns. ROI cross-classifications and cross-validations were performed in a similar fashion as above. One-sample *t*-tests (Bonferroni-corrected) showed across-domain classification was significantly above chance in dACC/pre-SMA (T_{21}=2.83, P=0.010) and precuneus (T_{21}=4.69, P=9.14×10^{-5}) but not in rlPFC (left: T_{21}=1.36, P=0.19; right: T_{21}=0.97, P=0.34), indicative of a generic confidence signal (Figure 3C, light colors). In contrast, within-domain cross-validation classification accuracy was significantly above chance in right rlPFC (T_{21}=3.74, P=0.001; Bonferroni-corrected). Importantly, classification accuracy in this ROI differed from the corresponding across-domain classification accuracy (paired *t*-test T_{21}=2.37, P=0.027; Figure 3C, dark colors; see Figure S7 for results split by domain). Together, these results suggest the coexistence of two kinds of CLR neural activity: dACC/pre-SMA and precuneus encode a generic confidence signal, whereas patterns of activity in right rlPFC were modulated by task, allowing within-domain classification of confidence level.

Searchlight analysis of confidence level-related activity. Here we leveraged the Follow trials as a control for low-level visuomotor confounds by exclusively masking out activity patterns associated with usage of the confidence scale (Figure S8 & Table S5). The remaining activity patterns can therefore be ascribed to confidence level-related signals that do not encode visual or motor features of the rating (Figure 3D). We found widespread across-domain classification of confidence level in a predominantly midline network including a large cluster encompassing dACC/pre-SMA, ventromedial PFC (vmPFC), and striatum (yellow). Domain-specific confidence activity was successfully decoded from right PFC (insula, IFG, BA 9, 46) in memory trials (blue) and was also independently decoded in both domains from dACC/pre-SMA.

Generalization of CLR-activity. We reasoned that even in the absence of confidence reports, confidence activity patterns may predict decision accuracy, a proxy of confidence on 'no-report' Follow trials (though we note that the neural correlates of confidence and performance are also likely to be partially distinct; 36,37). We trained a decoder with CLR-beta images on Confidence trials and tested it on beta images indexing accurate/inaccurate decisions on Follow trials as proxies of low/high confidence (and vice versa). This analysis confirmed that activity patterns in dACC/pre-SMA (T_{21}=2.578, P=0.0175) and right rlPFC (T_{21}=2.48, P=0.0215) could be used to predict accuracy levels in Follow trials above chance (Figure S10).

Metacognitive efficiency and confidence level-related activity classification. Finally, we reasoned that if confidence-related patterns of activation are contributing to metacognitive judgments, they should track individual differences in metacognitive efficiency. To test for such a relation, we asked whether individual metacognitive efficiency scores predicted searchlight clas-

sification accuracy of confidence level. Memory metacognitive efficiency predicted memory confidence classification accuracy in a cluster in right precuneus (cluster defining threshold $P<0.001$). This cluster was significant after small-volume correction for multiple comparisons at $P_{FWE}<0.05$ using an anatomical mask of bilateral precuneus (Figure 4). Perceptual metacognitive efficiency predicted perceptual confidence classification accuracy in bilateral vmPFC and cerebellum (cluster-defining threshold $P<0.001$, uncorrected). These results are in agreement with functional (10), anatomical (9) and lesion studies (19) that have linked individual variation in metacognitive accuracy to precuneus and PFC in memory and perception, respectively.

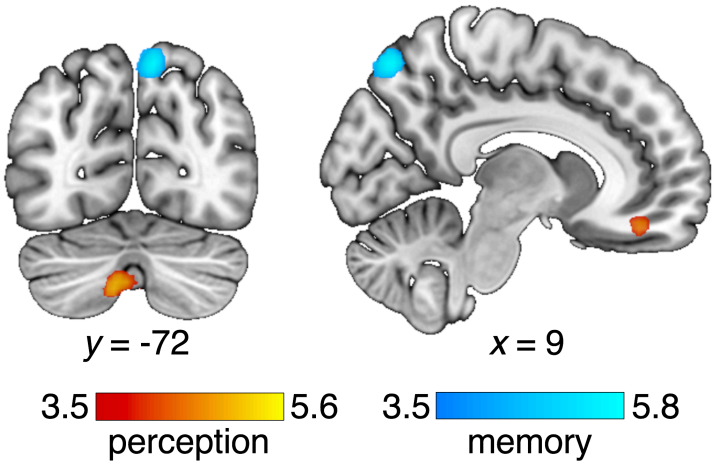

Figure 4. Metacognitive efficiency & confidence classification. Log M-ratio scores predicted within-domain classification accuracy of confidence level-related activity. Individual metacognitive efficiency in memory trials (blue) predicted classification accuracy in a cluster in right precuneus and perceptual metacognitive efficiency predicted confidence classification in perceptual trials (red) in bilateral ventral frontomedial areas and in the cerebellum. Images displayed at $P<0.001$, uncorrected. See main text for statistical analyses. Color bars indicate T-statistic. See Table S7.

Discussion

When performing a cognitive task, confidence estimates allow for comparisons of performance across a range of different scenarios (22). Such estimates must also carry information about the task context if they are to be used in decision-making. Here we investigated the domain-generality and domain-specificity of representations that support metacognition of perception and memory.

Unlike previous studies (9), subjects' performance was matched between domains for two different types of stimulus, thereby eliminating potential performance and stimulus confounds. Subjects' confidence ratings were also matched between domains and followed expected patterns of higher ratings after correct decisions than after incorrect decisions. Using MVPA, we

showed the existence of both domain-specific and domain-general metacognition-related activity during perceptual and memory tasks. We report three main findings, and discuss each of these in turn.

First, we obtained convergent evidence from both univariate and multivariate analyses that a cingulo-opercular network centered on dACC/pre-SMA encodes a generic signal predictive of confidence level across memory and perceptual tasks. Previous studies of metacognition have implicated the cingulo-opercular network in tracking confidence level (12,13,15,33,38). However, we go beyond these previous studies to provide evidence that these signals generalize to predict confidence across two distinct cognitive domains. This finding is consistent with posterior medial frontal cortex as a nexus for monitoring the fidelity of generic sensorimotor mappings, building on previous findings that error-related event-related potentials originating from this region are sensitive to variation in subjective certainty (39,40). The activity in this region was also consistently elevated by the requirement for a metacognitive judgment (12), although the pattern of these increases did not generalize across tasks.

Second, in anterior frontal regions we found activity patterns that tracked both the requirement for metacognitive judgments and level of confidence. Critically, however, confidence-related activity patterns were selective for domain in right rlPFC: they differed according to whether the subject was engaged in a metacognitive judgment about perception or memory. Such signals may support the "tagging" of confidence with contextual information, thereby facilitating the use of confidence for behavioral control (27,41). It is possible that anterior prefrontal regions combine generic confidence signals with domain-specific information to fine-tune decision-making and action selection in situations in which subjects need to regularly switch between tasks or strategies on the basis of their reliability (41). An alternative hypothesis, also compatible with our data, is that PFC first estimates the confidence level specifically for the current task, which is then relayed to medial areas to recruit the appropriate resources for cognitive control in a task-independent manner. Processing dynamics may also unfold simultaneously in both areas. These possibilities echo a longstanding debate in the cognitive control literature on the relative primacy of medial and lateral PFC in the hierarchy of control (42,43). Further inquiry and development of computational models of the hierarchical or parallel functional coupling of these networks in metacognition is necessary.

Third, we obtained convergent evidence that precuneus plays a specific role in metamemory judgments. In univariate fMRI analyses, we found the requirement for a metacognitive judgment recruited precuneus only on memory, but not perceptual, trials (Figure S5B). Individual metacognitive efficiency scores in memory trials predicted classification accuracy in precuneus (Figure 4), and individual differences in metacognitive efficiency scores in perceptual trials predicted classification accuracy in ventromedial prefrontal cortex (albeit at an uncorrected threshold). These findings are consistent with previous results that the medial parietal cortex makes a disproportional contribution to memory metacognition (9,14), and offer a potential explanation for a decrease in perceptual, but not memory, metacognitive efficiency seen in patients with frontal lesions (12). However, we do not wish to conclude that precuneus involvement is

specific to metamemory. We note that univariate negative correlations with confidence were found also on perceptual trials, and multivariate classification results in precuneus indicated the presence of both perceptual and memory-related signals. This dual involvement of the precuneus in perception and memory metacognition is consistent with previous studies which suggest a relationship between precuneus structure and visual perceptual metacognition (9,10).

Our experimental design assumes that visual perception and memory are distinct domains. We acknowledge that distinguishing between cognitive domains or individuating perceptual modalities is not straightforward (44). For instance, different modalities (e.g. vision, audition, touch, etc.) could be part of a unified perceptual domain for metacognitive purposes. Recent findings suggest metacognitive efficiency in one perceptual modality predicts metacognitive efficiency in others and that they share electrophysiological markers (29). However, it is an open question whether modality-specific patterns of metacognitive activity could be decoded using multivariate approaches. More research is needed on the neural architecture of metacognition in other cognitive domains, and whether this architecture changes in a graded or discrete fashion as a function of task or stimulus.

In summary, our results provide evidence for the co-existence of content-rich metacognitive representations in frontopolar cortex with generic confidence-related signals in frontoparietal and cingulo-opercular regions. Such an architecture may be appropriate for "tagging" lower-level feelings of confidence with higher-order contextual information to allow effective behavioral control. Our findings reconcile previously conflicting results on the domain-specificity/generality of metacognition, and lay the groundwork for a mechanistic understanding of reflective judgments of cognition.

MATERIALS AND METHODS

Participants

Thirty healthy subjects (ages 18-33, mean 24.97 years old; SD=4.44; 14 males) with normal or corrected-to-normal vision were monetarily compensated and gave written informed consent to participate in the study at the Center for Neural Science at New York University. The number of participants was determined a priori at n=30, which is in line with recent guidelines on neuroimaging sample sizes (45). The study protocols were approved by the local Institutional Review Board.

Task

Experiments were programmed in MATLAB (MathWorks) using Psychtoolbox (46). Prior to entering the scanner, participants were familiarized with the tasks and the confidence rating scale. During practice trials, subjects got feedback on their responses but did not rate confidence.

The brightness of a randomly located reference stimulus was fixed (mid-gray). The difference in brightness (Δb) between the two stimuli was titrated to calibrate the required brightness of the non-reference stimulus using a staircase procedure similar to previous experiments (10,12,19). We used a fixed large step size 2-down/1-up procedure where subjects practiced until they reached 15 reversals or 90 trials following recommended ratios (47) to target performance levels similar as the expected ones in the memory blocks. Subjects began the experiment with a Δb value determined by the average of the Δb values at each reversal, excluding the first one. Throughout the experiment, we kept a small step size staircase running to account for learning or tiredness. After computing independent thresholds for words and abstract shapes, subjects practiced one whole block (9 trials) in each condition that mimicked experimental blocks (72 trials total). Instructions emphasized that confidence ratings should reflect relative confidence and participants were encouraged to use all ratings.

The experiment consisted of 432 trials of two-alternative forced-choice perception and memory tasks divided into six scanner runs. Task domains were interleaved across runs and counterbalanced across subjects. Runs consisted of four alternations of Confidence and Follow blocks (nine trials each), in which subjects had to either rate their confidence by selecting a number from a 1-to-4 scale, or "follow the computer" by pressing the button corresponding to the highlighted number irrespective of their confidence. The highlighted number was yoked to the previous Confidence block ratings (randomized presentation order) to ensure subjects pressed the same buttons in both conditions during any given pair of blocks. Words and abstract shapes were used as stimuli (interleaved blocks, order counterbalanced across runs). Subjects were reminded at the beginning of each block of the condition, task, and stimulus type that would follow. Subjects used two fingers of their right hand to respond on an MRI-compatible button box: left stimulus (index) and right stimulus (middle). For confidence ratings, they used four fingers: 1 (index), 2 (middle), 3 (ring) and 4 (small). If subjects failed to provide either type of response within the allotted time (see Figure 1A for details), the trial was missed and an exclamation mark was displayed for the remainder of the trial. Failing to press the highlighted number counted as a missed trial. The whole experiment lasted ~1.5 hours.

Stimuli

Stimuli were presented on a black background. Throughout trials, before responses, a mid-gray fixation cross subtending 0.3 degrees of visual angle was presented between stimuli. The reference stimulus in the perceptual task and all stimuli in the memory task were mid-gray. All stimuli were surrounded by an isoluminant blue bounding box separated from the stimulus by a gap of at least 0.15 degrees of visual angle.

22- or 28-line abstract shapes were randomly created by specifying an (invisible) grid of 6×6 squares that subtended 4 degrees of visual angle where lines could connect two vertices horizontally, vertically, or diagonally. The first line always stemmed from the central vertex of the invisible grid randomly connecting one of the surrounding eight vertices to ensure shape centrali-

ty within the grid. The remaining lines were drawn sequentially, ensuring all lines were connected. Orientation and originating vertices were selected randomly.

All words were nouns of 6 to 12 letters with 1 to 4 syllables obtained from the Medical Research Council Psycholinguistic Database (48). In the perceptual task, words had high familiarity, concreteness, and imageability ratings (400-700). In the memory task, words had low ratings (100-400) to increase task difficulty. Each word and each shape was presented once throughout the experiment (across perceptual and memory blocks, including practice trials). All subjects were tested on the same words and shapes (counterbalanced across Confidence and Follow conditions across subjects). Words and rating scales were presented using DS-Digital font (40 points) to make their visual features similar to the abstract shapes.

To obtain stimulus sets of similar difficulty for shapes and words, we ran a series of pilot studies where participants rated abstract shapes' distinctiveness and then performed both tasks (171 Amazon Mechanical Turk participants [73 for shapes; 98 for words] and 6 subjects in the laboratory). Based on these results, we expected a mean performance of ~71% correct responses when 22- and 28-line distinctive shapes were used in the same block, and ~83% correct when long words (6-12 letters) with low concreteness, imageability and familiarity ratings (100-400) were used. To further increase difficulty, we created pairs of old and new words split between Confidence and Follow conditions (counterbalanced across subjects), blocked by similar semantic category (e.g. finance, argumentation, character traits, etc.), such that each new word within a block was freely associated with one old word (and when possible, vice versa) according to the University of South Florida free association normed database (49).

Behavioral data analysis

Data analysis was performed in MATLAB and statistical analysis in RStudio (50). We estimated metacognitive efficiency by computing log M-ratio (meta-d'/d'). d' is a signal detection theoretic measure of type 1 sensitivity, while meta-d' is a measure of type 2 sensitivity (i.e. the degree to which a subject discriminates correct form incorrect responses) expressed in the same units as type 1 sensitivity (d') (5,6). We used hierarchical Bayesian estimation to incorporate subject-level uncertainty in group-level parameter estimates (32). Certainty on this parameter was determined by computing the 95% high-density interval (HDI) from the posterior samples (51). Two subjects were discarded for missing more than 10% of the trials (i.e. >1 standard deviation from the average missed trials, which was 5%). Missed trials were not analyzed.

Multi-voxel pattern analysis

MVPAs were carried out in MATLAB using The Decoding Toolbox (35). We classified beta images from GLMs modeling JR- and CLR-activity in ROI and whole-brain searchlight analyses (see Supplemental Methods for fMRI data acquisition and preprocessing). ROI MVPAs were performed on normalized, smoothed images. Previous work has shown that these preprocessing steps have minimal impact on SVM classification accuracy, while allowing meaningful compari-

438 son across subject-specific differences in anatomy, as in standard fMRI analyses (52,53). A sin-
439 gle accuracy value per subject, per condition was extracted and used for group analysis.
440 Searchlight analyses used 12mm-radius spheres centered around a given voxel, for all
441 voxels, on spatially realigned and slice-timing corrected images from each subject to create
442 whole-brain accuracy maps. For group-level analyses, these individual searchlight maps were
443 spatially normalized and smoothed using a Gaussian kernel (8mm FWHM) and entered into one-
444 sample *t*-tests against chance accuracy (33,35). Whole-brain cluster inference was carried out in
445 the same manner as in univariate analysis (see Supplementary Methods).
446 Prior to decoding, for JR-activity classification we modeled the rating periods in the Con-
447 fidence and Follow conditions in two regressors of interest per run. For CLR-activity, we col-
448 lapsed ratings 1 and 2 into a low confidence regressor and ratings 3 and 4 into a high confidence
449 regressor to allow binary classification. Despite being encouraged to use the four levels of confi-
450 dence in every run, two subjects did not provide ratings for one of the classes in at least one run
451 and were left out from the analysis to avoid entering unbalanced training data into the classifier.
452 The main result of the ROI analysis (Figure 3C) is not affected when the unbalanced data of the-
453 se two subjects are included: within-domain classification accuracy in right rlPFC (M=62.22%)
454 is significantly above chance (T_{23}= 4.22; P=0.0003) and different from across-domain classifica-
455 tion accuracy (M=52.84%; paired *t*-test: T_{23}= 2.54; P=0.018). The remaining parameters were
456 specified as in the univariate case (see Supplementary Methods).
457 In independent across-domain classifications, we used the run-wise beta images reflect-
458 ing JR- and CLR-activity as pattern vectors in a linear support vector classification model (as
459 implemented in LIBSVM). We assigned each vector from each domain a label corresponding to
460 Confidence (1) and Follow (-1) in the JR-analysis and Low Confidence (-1) and High Confi-
461 dence (1) in CLR-analysis. We trained a support vector machine (SVM) with the vectors from
462 one domain (3 per class, 6 in total) and tested the decoder on the 6 vectors from the other domain
463 (and vice versa). We report the average classification accuracies of these two-way cross-
464 classifications.
465 For within-domain classifications, we ran independent leave-one-run-out cross-
466 validations for each domain on JR-activity (Confidence vs Follow) and CLR-activity (Low vs
467 High confidence). The pattern vectors from two of the three runs in each domain were used to
468 train an SVM to predict the same classes in the vectors from the left-out run. We compared the
469 true labels of the left-out run with the labels predicted by the model and iterated this process for
470 the other two runs to calculate a mean cross-validated accuracy independently for each domain.
471 Metacognitive scores (log M-ratio) were estimated independently for perceptual and
472 memory blocks as well as a single score collapsing both domains. These scores were inserted as
473 a covariate in second level analyses for within- and across-domain classifications of confidence
474 level-related activity. The covariate contrast was then obtained for each domain map and for the
475 across-domain results. An anatomical map of the precuneal region was extracted from the Har-
476 vard-Oxford probabilistic atlas of human cortical brain areas, and used for performing a small-
477 volume correction for the within-memory results.

Acknowledgements

We thank Aurelio Cortese for helping with the MVPA design. This work was partially supported by the National Institute of Neurological Disorders and Stroke of the National Institutes of Health (Grant No. R01NS088628) to H.L. and S.M.F., by the Templeton Foundation (Grant No. 21569) to H.L., and by a Sir Henry Wellcome Fellowship to S.M.F. (Grant No. WT096185).

References

1. Metcalfe J, Shimamura AP, editors. Metacognition. Cambridge, MA: MIT Press; 1994.
2. Flavell JH. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. American Psychologist. American Psychological Association; 1979;34(10):906–11.
3. Nelson TO, Narens L. Metamemory: A Theoretical Framework and New Findings. In: Bower G, editor. The Psychology of Learning and Motivation. New York: Academic Press; 1990. pp. 125–73.
4. Fleming SM, Dolan RJ, Frith CD. Metacognition: computation, biology and function. Phil Trans R Soc B. The Royal Society; 2012;367(1594):1280–6.
5. Maniscalco B, Lau H. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. Conscious Cogn. 2012;21:422–30.
6. Fleming SM, Lau H. How to measure metacognition. Front Hum Neurosci. 2014 Jul;8.
7. Galvin SJ, Podd JV, Drga V, Whitmore J. Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. Psychon Bull Rev. Springer; 2003;10(4):843–76.
8. Overgaard M, Sandberg K. Kinds of access: different methods for report reveal different kinds of metacognitive access. Phil Trans R Soc B. 2012 Apr 9;367(1594):1287–96.
9. McCurdy LY, Maniscalco B, Metcalfe J, Liu KY, de Lange FP, Lau H. Anatomical Coupling between Distinct Metacognitive Systems for Memory and Visual Perception. J Neurosci. 2013 Jan 30;33(5):1897–906.
10. Fleming SM, Weil RS, Nagy Z, Dolan RJ, Rees G. Relating Introspective Accuracy to Individual Differences in Brain Structure. Science. 2010 Sep 16;329(5998):1541–3.
11. Maniscalco B, McCurdy LY, Odegaard B, Lau H. Limited Cognitive Resources Explain a Trade-Off between Perceptual and Metacognitive Vigilance. J Neurosci. 2017 Feb 1;37(5):1213–24.
12. Fleming SM, Huijgen J, Dolan RJ. Prefrontal Contributions to Metacognition in Perceptual Decision Making. J Neurosci. 2012 May 2;32(18):6117–25.
13. Fleck MS, Daselaar SM, Dobbins IG, Cabeza R. Role of Prefrontal and Anterior Cingulate Regions in Decision-Making Processes Shared by Memory and Nonmemory Tasks. Cereb Cortex. 2006;16(11):1623–30.
14. Baird B, Smallwood J, Gorgolewski KJ, Margulies DS. Medial and Lateral Networks in Anterior Prefrontal Cortex Support Metacognitive Ability for Memory and Perception. J

518 Neurosci. 2013 Oct 16;33(42):16657–65.
519 15. Hilgenstock R, Weiss T, Witte OW. You'd Better Think Twice: Post-Decision Perceptual
520 Confidence. Neuroimage. 2014 Oct 1;99:323–31.
521 16. Yokoyama O, Miura N, Watanabe J, Takemoto A, Uchida S, Sugiura M, et al. Right fron-
522 topolar cortex activity correlates with reliability of retrospective rating of confidence in
523 short-term recognition memory performance. Neurosci Res. 2010 Nov;68(3):199–206.
524 17. Shimamura AP, Squire LR. Memory and metamemory: a study of the feeling-of-knowing
525 phenomenon in amnesic patients. J Exp Psychol Learn. 1986 Jul;12(3):452–60.
526 18. Schnyer DM, Verfaellie M, Alexander MP, LaFleche G, Nicholls L, Kaszniak AW. A role
527 for right medial prefrontal cortex in accurate feeling-of-knowing judgments: evidence
528 from patients with lesions to frontal cortex. Neuropsychologia. 2004 Jan;42(7):957–66.
529 19. Fleming SM, Ryu J, Golfinos JG, Blackmon KE. Domain-specific impairment in meta-
530 cognitive accuracy following anterior prefrontal lesions. Brain. 2014 Aug
531 6;137(10):2811–22.
532 20. Pouget A, Drugowitsch J, Kepecs A. Confidence and certainty: distinct probabilistic quan-
533 tities for different goals. Nature Neurosci. 2016 Feb 23;19(3):366–74.
534 21. Fleming SM, Daw ND. Self-evaluation of decision-making: A general Bayesian frame-
535 work for metacognitive computation. Psychol Rev. 2017;124(1):91–114.
536 22. de Gardelle V, Mamassian P. Does Confidence Use a Common Currency Across Two
537 Visual Tasks? Psychol Sci. 2014 Apr 3.
538 23. Cole MW, Reynolds JR, Power JD, Repovs G, Anticevic A, Braver TS. Multi-task con-
539 nectivity reveals flexible hubs for adaptive task control. Nature Neurosci. Nature Re-
540 search; 2013 Sep 1;16(9):1348–55.
541 24. Heekeren HR, Marrett S, Ruff DA, Bandettini PA, Ungerleider LG. Involvement of hu-
542 man left dorsolateral prefrontal cortex in perceptual decision making is independent of re-
543 sponse modality. Proc Natl Acad Sci USA. 2006 Jun 27;103(26):10023–8.
544 25. Holroyd CB, Yeung N, Coles MGH, Cohen JD. A Mechanism for Error Detection in
545 Speeded Response Time Tasks. J Exp Psychol Gen. 2005;134(2):163–91.
546 26. Gehring WJ, Goss B, Coles MGH, Meyer DE, Donchin E. A Neural System for Error De-
547 tection and Compensation. Psychol Sci. Sage Publications, Inc; 1993 Nov 1;4(6):385–90.
548 27. Donoso M, Collins AGE, Koechlin E. Foundations of human reasoning in the prefrontal
549 cortex. Science. 2014 Jun 27;344(6191):1481–6.
550 28. Ais J, Zylberberg A, Barttfeld P, Sigman M. Individual consistency in the accuracy and
551 distribution of confidence judgments. Cognition. 2016 Jan;146:377–86.
552 29. Faivre N, Filevich E, Solovey G, Kuhn S, Blanke O. Behavioural, modeling, and electro-
553 physiological evidence for domain-generality in human metacognition. bioRxiv. 2016.
554 30. Kelemen WL, Frost PJ, Weaver CA. Individual differences in metacognition: evidence
555 against a general metacognitive ability. Mem Cognit. 2000 Jan;28(1):92–107.
556 31. Vo VA, Li R, Kornell N, Pouget A, Cantlon JF. Young Children Bet on Their Numerical
557 Skills: Metacognition in the Numerical Domain. Psychol Sci. 2014 Sep 10;25(9):1712–21.

558 32. Fleming SM. Hierarchical Bayesian estimation of metacognitive efficiency from confi-
559 dence ratings. Neurosci Consc. 2017 Apr 22;3(1).
560 33. Hebart MN, Schriever Y, Donner TH, Haynes JD. The Relationship between Perceptual
561 Decision Variables and Confidence in the Human Brain. Cereb Cortex. 2014;26(1):118–
562 30.
563 34. Cole MW, Ito T, Braver TS. The Behavioral Relevance of Task Information in Human
564 Prefrontal Cortex. Cereb Cortex. 2016 May 17;26(6):2497–505.
565 35. Hebart MN, Görgen K, Haynes JD. The Decoding Toolbox (TDT): a versatile software
566 package for multivariate analyses of functional imaging data. Front Neuroinform. 2015
567 Jan;8.
568 36. Rounis E, Maniscalco B, Rothwell JC, Passingham RE, Lau H. Theta-burst transcranial
569 magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness.
570 Cogn Neurosci. 2010 Aug 18;1(3):165–75.
571 37. Cortese A, Amano K, Koizumi A, Kawato M, Lau H. Multivoxel neurofeedback selec-
572 tively modulates confidence without changing perceptual performance. Nat Commun.
573 2016 Dec 7;7.
574 38. De Martino B, Fleming SM, Garrett N, Dolan RJ. Confidence in value-based choice. Na-
575 ture Neurosci. 2013;16:105–10.
576 39. Boldt A, Yeung N. Shared Neural Markers of Decision Confidence and Error Detection. J
577 Neurosci. 2015 Feb 25;35(8):3478–84.
578 40. Scheffers MK, Coles M. Performance monitoring in a confusing world: error-related brain
579 activity, judgments of response accuracy, and types of errors. J Exp Psychol Human.
580 2000;26(1):141–51.
581 41. Purcell BA, Kiani R. Hierarchical decision processes that operate over distinct timescales
582 underlie choice and changes in strategy. Proc Natl Acad Sci USA. 2016 Aug
583 2;113(31):E4531–40.
584 42. Tang H, Yu H-Y, Chou C-C, Crone NE, Madsen JR, Anderson WS, et al. Cascade of neu-
585 ral processing orchestrates cognitive control in human frontal cortex. eLife. 2016 Feb
586 18;5.
587 43. Kerns JG, Cohen JD, Stenger VA, Carter CS. Prefrontal cortex guides context-appropriate
588 responding during language production. Neuron. 2004 Jul 22;43(2):283–91.
589 44. Macpherson F. Introduction: Individuating the Senses. In: Macpherson F, editor. The
590 Senses. Oxford: Oxford University Press; 2011. pp. 3–42.
591 45. Poldrack RA, Baker CI, Durnez J, Gorgolewski KJ, Matthews PM, Munafò MR, et al.
592 Scanning the horizon: towards transparent and reproducible neuroimaging research. Nat
593 Rev Neurosci. 2017 Feb;18(2):115–26.
594 46. Brainard DH. The Psychophysics Toolbox. Spat Vis. 1997 Jan 1;10(4):433–6.
595 47. García-Pérez MA. Forced-choice staircases with fixed step sizes: asymptotic and small-
596 sample properties. Vision Res. 1998 Jun;38(12):1861–81.
597 48. Wilson M. MRC Psycholinguistic Database: Machine-usable dictionary, version 2.00. Be-

hav Res Methods Instrum Comput. 1988;20(1):6–10.
49. Nelson DL, McEvoy CL, Schreiber TA. The University of South Florida free association, rhyme, and word fragment norms. Behav Res Methods Instrum Comput. 2004 Aug;36(3):402–7.
50. R Studio Team. RStudio: Integrated Development for R. Boston, MA: RStudio, Inc; 2015. Available from: http://www.rstudio.com
51. Kruschke J. Doing Bayesian Data Analysis. Academic Press; 2010.
52. Kamitani Y, Sawahata Y. Spatial smoothing hurts localization but not information: Pitfalls for brain mappers. Neuroimage. 2010 Feb 1;49(3):1949–52.
53. Op de Beeck HP. Against hyperacuity in brain reading: spatial smoothing does not hurt multivariate fMRI analyses? Neuroimage. 2010 Feb 1;49(3):1943–8.

638 **Supplementary Information**
639
640 ***Results from univariate analysis of fMRI data***
641
642 *Judgment-related activity*. In a Confidence>Follow contrast (Figure S1A), we found significant
643 activations across domains in dorsal anterior cingulate cortex (dACC; BA 24 & 32), pre-
644 supplementary motor area (pre-SMA), bilateral insulae, superior and middle frontal gyri (SFG &
645 MFG), and lingual gyrus, consistent with previous findings (1).
646
647 *Confidence level-related activity*. We found positive correlations between confidence level and
648 activation in left primary motor cortex and SMA, middle and anterior cingulate cortex (BA 11,
649 25), as well as in the left hippocampus, and right cerebellum and lingual gyrus (Figure S1B, hot
650 colors). Negative correlations with confidence were found in dACC/ pre-SMA, bilateral middle
651 and inferior frontal gyrus (BA 9, 46), bilateral parietal cortex (BA 40), and left cerebellum and
652 lingual gyrus (Figure S1B, cool colors). Similar negative correlations between confidence and
653 activation in dACC/pre-SMA, parietal cortex and PFC have been reported in several previous
654 studies (1-4).
655
656 *Confidence level-related activity by domain*. We further assessed these data splitting them by
657 domain. Positive correlations with confidence in perception and memory trials were shared in
658 pre- and post-central gyri, consistent with subjects using different fingers from their right hand to
659 rate confidence (green in Figure S2A). Shared negative correlations with confidence were found
660 in right dorsal PFC (green in Figure S2B). In perceptual trials, a negative correlation with confi-
661 dence was significant in dACC/pre-SMA, bilateral parietal regions and left inferior and middle
662 frontal gyri (BA9, 10, 46), whereas in memory trials the same negative correlation was found in
663 lingual gyrus and right superior frontal gyrus (BA6).
664
665 *ROI analysis.* To separately quantify these effects for each task domain, we focused on four a
666 priori regions of interest (ROI) in dACC/pre-SMA, bilateral rostrolateral prefrontal cortex
667 (rlPFC) and precuneus (PCUN) (Figure S3A), which previous studies have found to be recruited
668 by perceptual and memory metacognition (1,2,4-6). An analysis of confidence level-related ac-
669 tivity recapitulated the whole-brain results, with negative relationships between confidence and
670 activity in dACC/pre-SMA. Importantly, no significant differences in beta values were found
671 between domains in any of the ROIs (all paired *t*-tests: $T_{23}<1.6$; $P>0.05$; Figure 4). Unlike the
672 whole-brain results by domain, the ROI analysis was suggestive of a lack of domain-specificity
673 in confidence level-related activity. This ambiguity suggests a potential insensitivity of univari-
674 ate analyses to putative domain-specific differences in confidence encoding.
675 With respect to judgment-related activity, in a series of repeated measures 2×2 ANOVAS
676 (CONDITION × TASK) we found greater activity on Confidence compared to Follow trials in both
677 tasks in all ROIs except precuneus, where this increase was observed only for memory trials

(Figure S3B). These results are compatible with previous findings indicating a distinctive role of precuneus during memory metacognition (4,6).

Supplementary Methods

fMRI data acquisition

Brain images were acquired using a 3T Allegra scanner (Siemens). BOLD-sensitive functional images were acquired using a T2*-weighted gradient-echo echo-planar images (42 transverse slices, interleaved acquisition; TR, 2.34s; TE, 30ms; matrix size: 64×64; 3×3mm in-plane resolution; slice thickness: 3mm; flip angle: 90°; FOV: 126mm). The main experiment consisted of three runs of 210 volumes and three runs of 296 volumes for the perceptual and memory tasks, respectively. We collected a T1-weighted MPRAGE anatomical scan (1×1×1mm voxels; 176 slices) and local field maps for each subject.

fMRI data preprocessing

Imaging analysis was carried out using SPM12 (Statistical Parametric Mapping; www.fil.ion.ucl.ac.uk/spm). The first five volumes of each run were discarded to allow for T1 stabilization. Functional images were realigned and unwarped using local field maps (7) and then slice-timing corrected (8). Each participant's structural image was segmented into gray matter, white matter, cerebral spinal fluid, bone, soft tissue, and air/background images using a nonlinear deformation field to map it onto template tissue probability maps (9). This mapping was applied to both structural and functional images to create normalized images to Montreal Neurological Institute (MNI) space. Normalized images were spatially smoothed using a Gaussian kernel (8mm FWHM). We set a within-run 1mm rotation and 4mm affine motion cut-off criterion, which led to the exclusion of 4 subjects, leaving a total of 24 subjects whose functional and behavioral data were fully analyzed.

Univariate analysis

All our GLMs focus on the rating period (from initial response to confidence rating or number following) and motion correction parameters were entered as covariates of no interest along with a constant term per run. Regressors were convolved with a canonical hemodynamic response function (HRF). Low-frequency drifts were excluded with a 1/128Hz high-pass filter. Missed trials were not modeled. For JR-analyses, we created a GLM with two regressors of interest per run to estimate BOLD response amplitudes in each voxel during the rating period in each trial of the Confidence and Follow blocks. For the CLR-parametric modulation analysis, a GLM was used to estimate BOLD responses in Confidence blocks. There were two regressors of interest in each run, one modeling the confidence rating period and another that encoded a parametric modulation by confidence rating.

Statistical inference

For the JR-analysis, single-subject contrast images of the Confidence and Follow regressors were entered into a second-level random effects analysis using one-sample *t*-tests against zero to assess group-level significance. For the CLR-parametric modulation analysis, single-subject contrast images of the parametric modulator were entered into a similar second-level random effects analysis. Activations were visualized using Surf Ice (https://www.nitrc.org/projects/surfice/) and MRIcro (http://www.mccauslandcenter.sc.edu/crnl/mricro).

ROI analysis

To define regions of interest, 12mm spheres were centered at MNI coordinates identified from previous literature (Figure S4). Beta values were extracted from subjects' Confidence>Follow contrast and confidence parametric modulator contrast images for the JR and CLR analyses, respectively.

Supplementary Figures

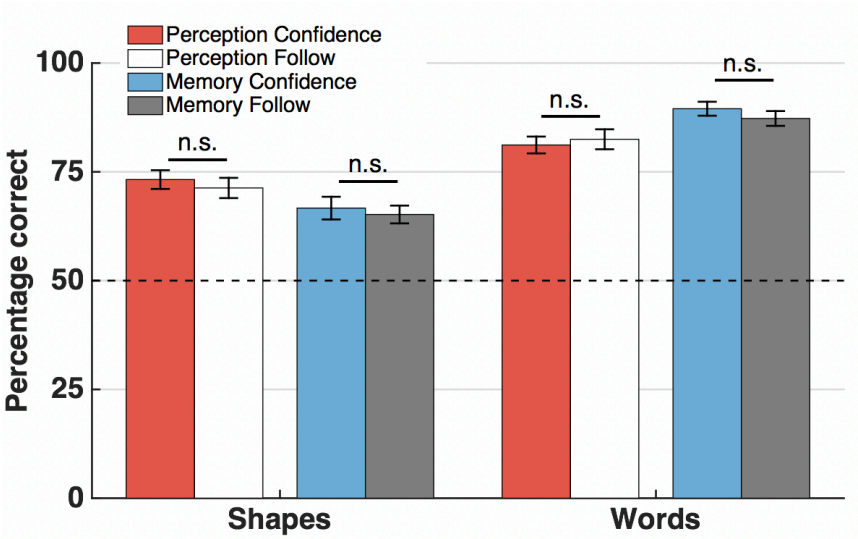

Figure S1. Task performance (percentage correct) was matched across Confidence and Follow conditions for each combination of task and stimulus type. In a 2×2×2 repeated measures ANOVA (CONFIDENCE/FOLLOW × PERCEPTION/MEMORY × SHAPES/WORDS), there was no three-way interaction ($F_{1,23}$=2.573, P=0.12) and no effect of rating condition ($F_{1,23}$=3.036, P=0.095). None of the four paired t-tests comparing performance between the Confidence and Follow conditions returned a significant difference (P>0.05).

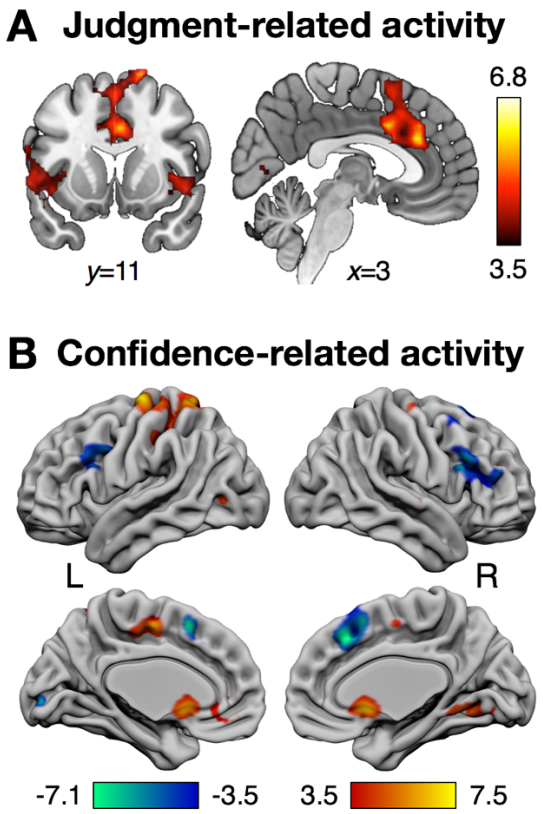

Figure S2. fMRI univariate analysis results. (A) Activity from the Confidence>Follow contrast. **(B)** Activity parametrically modulated by level of reported confidence. Hot colors indicate a positive correlation with confidence and cool colors a negative correlation. All displayed activations are significant at cluster-defining threshold $P<0.001$, corrected for multiple comparisons at $P_{FWE}<0.05$. Images displayed at $P< 0.001$, uncorrected. Color bars reflect T-statistic (see Tables S1 & S2). L=left; R=right.

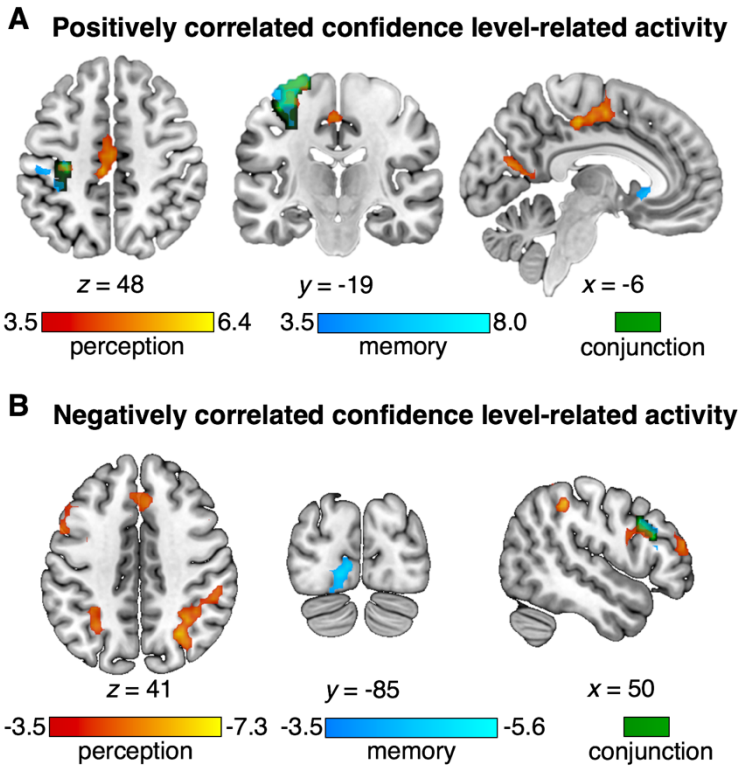

Figure S3. Confidence level-related fMRI univariate analysis by domain. (A) Activity positively correlated by level of reported confidence in perceptual (red) and memory (trials). Green areas indicate significant voxels in both domains. Common regions were active only in left sensorimotor areas. **(B)** Activity negatively correlated by level of reported confidence; colors indicate same as in (A). Areas common to both domains were found in right dorsolateral PFC. Confidence was anti-correlated in perception with bilateral parietal regions, dACC/pre-SMA, and left dorsolateral PFC; in memory, left lingual gyrus was active. All displayed activations are significant at cluster-defining threshold $P<0.001$, corrected for multiple comparisons $P_{FWE}<0.05$. Images displayed at $P<0.001$, uncorrected. Graded color bars reflect T-statistic (see Table S3).

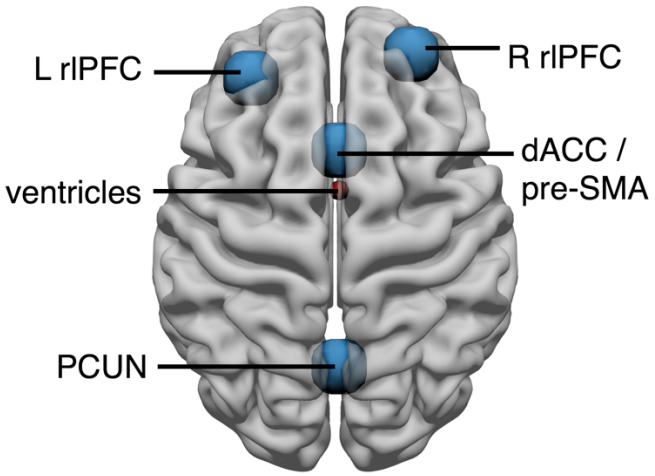

Figure S4. Regions of Interest. We created 12mm-radius spherical binary masks for four regions of interest (ROI) based on previously reported peak MNI coordinates. ROIs in left rostrolateral prefrontal cortex (L rlPFC) [-33, 44, 28], right rlPFC (R rlPFC) [27, 53, 25] and dorsal anterior cingulate cortex/pre-supplementary motor area (dACC/pre-SMA) [0, 17, 46] were created based on (1). The mask for precuneus (PCUN) [0, -64, 24] was based on (6). The MNI x-coordinates for the dACC/pre-SMA and PCUN masks were set to 0 to ensure bilaterality. As a control, we also created a 6mm-radius sphere centered at the ventricles [0 2 15].

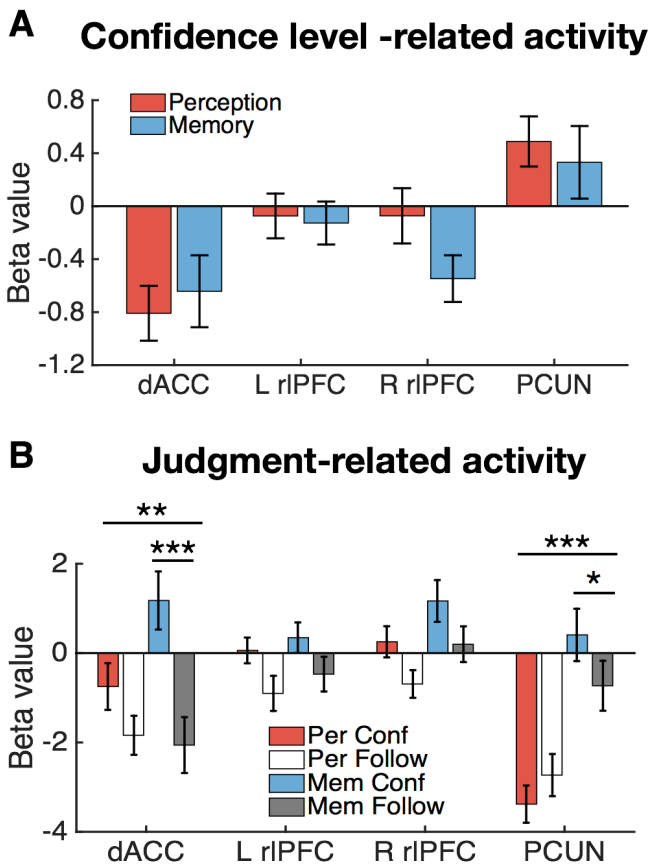

Figure S5. Univariate ROI analysis results. (A) Estimated mean beta values of confidence level-related activity. No significant difference between domains was found in any ROI. **(B)** Estimated mean beta values of judgment-related activity. Repeated measures 2×2 ANOVA (CONDITION × TASK) on each ROI revealed that in dACC/pre-SMA there was a greater activity on Confidence compared to Follow trials ($F_{1,23}$=19.34, P=0.0002) and this difference was greatest on memory trials (interaction: $F_{1,23}$=12.16, P=0.0019; post hoc paired t-test, T_{23}=5.472, P=0.0001). In frontopolar cortex there was a consistently heightened activity on Confidence compared to Follow trials in both left rlPFC ($F_{1,23}$=6.62, P=0.017) and right rlPFC ($F_{1,23}$=9.282, P=0.0057). Finally, in precuneus, activity was overall greater on memory compared to perception trials ($F_{1,23}$=36.78, P=0.000003), and differentiated between Confidence and Follow trials in the memory task (interaction: $F_{1,23}$=15.86, P=0.0006; paired t-test, T_{23}=2.431, P=0.023). Error bars indicate s.e.m. *** P<0.001 ** P<0.01 *P<0.05

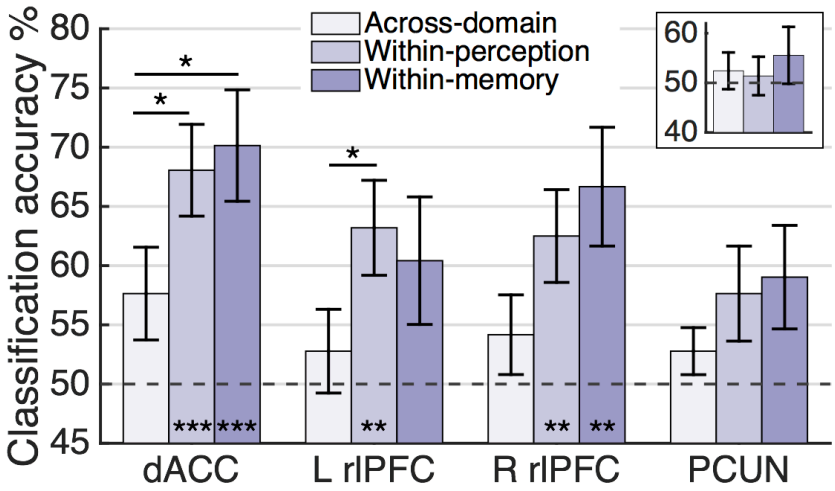

Figure S6. Judgment-related MVPA ROI results split by domain. ROI results for across-domain (left bars; same as in Figure 5A and presented here for comparison) and mean within-perception (middle bars) and within-memory (right bars) classification accuracy of judgement-related activity. In dACC/pre-SMA both within-perception (T_{23}=4.66, P=0.0001) and within-memory (T_{23}=4.28; P=0.0003) classifications were different from chance, and each was significantly different from across-domain classification accuracy (paired t-tests perception: P=0.0261; memory: P=0.026). In left rlPFC, classification accuracy within-perception was different from chance (T_{23}=3.29; P=0.003) and from across-domain classification accuracy (paired t-test: P=0.043). In right rlPFC within-perception (T_{23}=3.19; P=0.004) and within-memory (T_{23}=3.32; P= 0.003) classifications were significantly above chance. All reported t-tests against chance are significant after Bonferroni correction (α=0.05/4=0.0125). Inset: Classification accuracy in the ventricles, tested as a control, was not different from chance (perception: P=0.72; memory: P=0.34). *** $P \leq 0.001$ ** $P \leq 0.01$ * $P < 0.05$

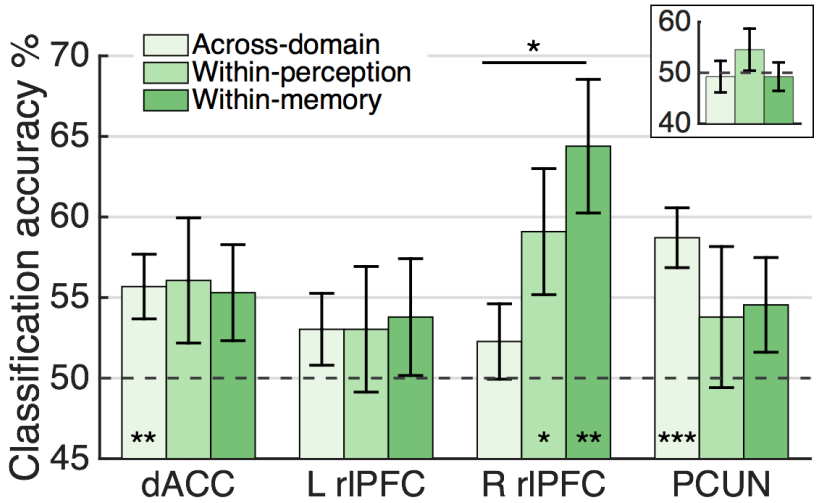

Figure S7. Confidence level-related MVPA ROI results split by domain. ROI results for across-domain (left bars; same as in Figure 5C and presented here for comparison) and mean within-perception (middle bars) and within-memory (right bars) classification accuracy of confidence level-related activity. In right rlPFC, within-memory (T_{21}=3.47; P=0.002) classification accuracy was significantly above chance (Bonferroni corrected). It was also different from across-domain classification accuracy (paired t-test T_{21}=2.39; P=0.026). Within-perception classification accuracy was significantly above chance, uncorrected (T_{21}=2.32; P=0.03). Inset: Classification accuracy in the ventricles was not different from chance (perception: P=0.28; memory: P=0.79). *** $P \leq 0.001$ ** $P \leq 0.01$ * $P < 0.05$

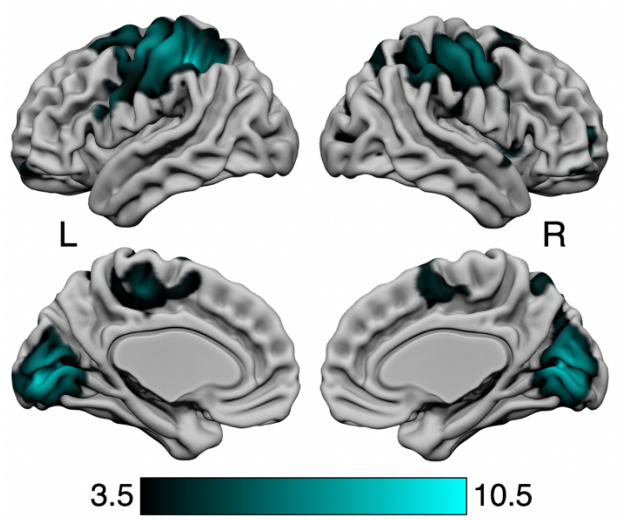

Figure S8. Low-level visuomotor features of the rating scale. We trained a classifier with subjects' low (1-2) and high (3-4) ratings from the Confidence condition and tested on trials from the Follow condition (and vice versa). Since rating-scale numbers in Follow trials were not related to confidence level, any accurately classified voxel could be expected to reflect visuomotor requirements of using the confidence scale rather than evaluating confidence. Accordingly, this analysis identified regions of visual cortex, parietal cortex and primary motor areas. This group-level accuracy map was used as an exclusive mask in whole-brain searchlight analyses of CLR-activity to eliminate visuomotor-related activations. All clusters are significant at $P_{FWE}<0.05$ cluster-level corrected for multiple comparisons using a cluster-defining threshold of $P<0.001$. Image displayed at $P<0.001$, uncorrected. The bar indicates T-score (Table S5).

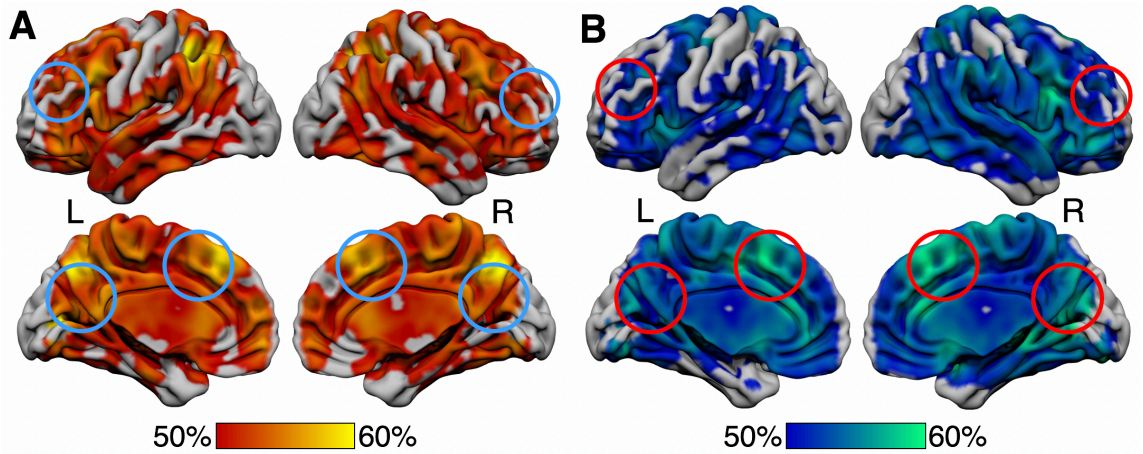

Figure S9. Mean raw confidence level-related classification accuracy scores. For illustration purposes, we present here whole-brain maps of group mean raw accuracy scores for within-domain confidence level-related classifications (Figure 5D), exclusively masked by visuomotor activity (Figure S8). Color bars indicate mean classification accuracy. Circles indicate the four selected ROIs (Figure 4 – figure supplement 1). **(A)** Within-perception. **(B)** Within-memory.

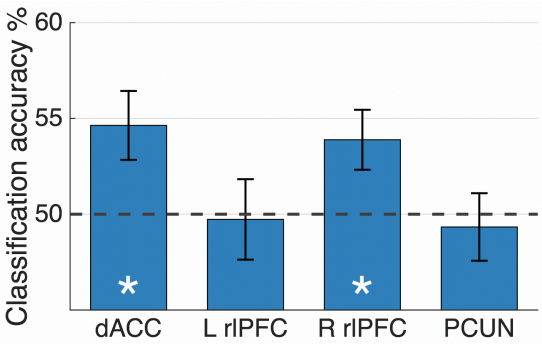

Figure S10. Classification accuracy of confidence level-related activity using accuracy as proxy. Cross-classification accuracy results of training a decoder on CLR-activity and testing it on activity from the Follow condition using accurate vs inaccurate trials as proxies of Low and High confidence (and vice versa). Classification accuracy was above chance in dACC/pre-SMA (T_{21}=2.58, P=0.0175) and right rlPFC (T_{21}=2.48, P=0.0215). Cross-classification of activity in left rlPFC (T_{21}=-0.13, P=0.898) and precuneus (T_{21}=-0.38, P= 0.71) was at chance. Bars indicate group means and bars s.e.m. * P<0.05

836 **Table S1. Judgment-related univariate activity for Confidence>Follow contrast.** Significant
837 activations at cluster-defining threshold $P<0.001$, corrected for multiple comparisons at
838 $P_{FWE}<0.05$
839

Label	**Voxels at *P*<0.001**	**P_{FWE} cluster-corrected**	**Peak *z*-score**	**Peak voxel MNI coordinates**	**Laterality**
Anterior cingulate gyrus, dACC, pre-SMA (BA 6, 24,32)	802	<0.001	4.97	6, 29, 26	L/R
Insula	323	<0.001	4.31	-27, 17, 8	L
Superior and middle frontal gyri	75	0.018	4.07	30, 41, 17	R
Cuneus, lingual gyrus	86	0.010	3.97	9, -67, -10	R
Insula, superior temporal gyrus	117	0.002	3.80	39, 11, 2	R

840

841 **Table S2. Confidence level-related univariate activity parametric correlations.** Significant
842 activations at cluster-defining threshold $P<0.001$, corrected for multiple comparisons at
843 $P_{FWE}<0.05$
844

Correlation	Label	Voxels at $P<0.001$	P_{FWE} cluster-corrected	Peak *z*-score	Peak voxel MNI coordinates	Laterality
Positive	Post- and pre-central gyri	848	<0.001	5.30	-33, -16, 71	L
	BA 11, 25	445	<0.001	5.02	-6, 17, -4	L/R
	Lingual gyrus, cerebellum	323	<0.001	4.85	12, -58, -10	R
	SMA, BA 31	228	<0.001	4.82	-3, -10, 53	L/R
	Hippocampus, parahipocampal gyrus	59	0.019	4.43	-27, -16, -16	L
	Post- and pre-central gyri	107	0.001	4.41	42, -22, 47	R
	Cuneus	53	0.030	4.12	18, -82, 35	R
	Posterior cingulate	51	0.035	3.81	-6, -55, 14	L
Negative	dACC, pre-SMA	553	<0.001	5.12	6, 29, 44	L/R
	Cerebellum, lingual gyrus, cuneus	287	<0.001	4.85	-39, -61, -43	L
	Middle and inferior frontal gyrus, BA 9	272	<0.001	4.55	-42, 17, 26	L
	Middle and inferior frontal gyrus, BA 46, 9	498	<0.001	4.39	48, 17, 29	R
	Parietal cortex, BA 40	135	<0.001	3.97	36, -52, 44	R
	Parietal cortex	71	0.008	3.81	-30, -61, 47	L

845
846

847 **Table S3. Confidence level-related univariate activity parametric correlations by domain.**
848 Significant activations at cluster-defining threshold $P<0.001$, corrected for multiple comparisons
849 at $P_{FWE}<0.05$
850

Correlation	Label	Voxels at $P<0.001$	P_{FWE} cluster-corrected	Peak z-score	Peak voxel MNI coordinates	Laterality
Positive						
Perception	Precentral gyrus, BA6	231	<0.001	4.78	-30, -25, 53	L
	Paracentral lobule (supplementary motor area), BA6	184	<0.001	4.55	-3, -16, 50	L/R
	Precuneus, posterior cingulate	139	<0.001	4.30	15, -73, 20	L/R
Memory	Precentral & postcentral gyri, BA6	723	<0.001	5.46	-30, -16, 71	L
	Cerebellum	55	0.037	4.54	-15, -37,-13	L
	Cerebellum	65	0.018	4.41	24, -46, -25	R
	Precuneus	63	0.021	4.40	21, -49, -17	R
	Middle Temporal Gyrus	52	0.045	4.08	-45, -73, 2	L
	Anterior cingulate, BA25	61	0.024	4.05	-6, 20, 1	L/R
Conjunction	Precentral & postcentral gyri, BA6	167			-30, -25, 53; -30, -16, 71	L
Negative						
Perception	Inferior parietal lobule, precuneus, BA7, BA40	434	<0.001	5.22	33, -61, 32	R
	Inferior & middle frontal gyri, BA9, BA10, BA46	364	<0.001	4.46	48, 5, 23	R
	Inferior & middle frontal gyri, BA9	214	<0.001	4.36	-45, 5, 29	L
	dACC, pre-SMA, BA8	239	<0.001	4.25	6, 29, 44	L/R
	Superior and inferior parietal lobules, precuneus, BA7, BA40	121	<0.001	4.16	-30, -61, 47	L
Memory	Lingual gyrus, cuneus	106	0.001	4.42	-6, -88, 2	L
	Superior & middle frontal gyri, BA6	94	0.003	4.04	36, 14, 56	R
	Inferior & middle frontal gyri	87	0.004	3.89	51, 17, 32	R
Conjunction	Middle frontal gyrus, BA9	29			51, 17, 32	R

851

852 **Table S4. Judgment-related activity in whole-brain searchlight classification accuracy map.**
853 Significant activations at cluster-defining threshold $P<0.001$, corrected for multiple comparisons
854 at $P_{FWE}<0.05$
855

Classification	Label	Voxels at $P<0.001$	P_{FWE} cluster-corrected	Peak *z*-score	Peak voxel MNI coordinates	Laterality
Across-domain	Insula/IFG	592	0.001	4.31	-33, 23, 5	L
	SFG & BA9 dACC/pre-SMA	928	<0.001	4.29 4.19 4.01	24, 53, 38 -12, 14, 41 12, 44, 23	R L R
	Supramarginal gyrus, BA 40	207	0.039	3.93	-51, -55, 29	L
	IFG/Insula/STG	275	0.016	3.61	48, 5, -13	R
Within-perception	Superior, Middle, Medial, Inferior FG, dACC, pre-SMA, BA 8, 10, 32	4699	<0.001	5.10	-12, 14, 59	L/R
	Parietal cortex, precuneus, supramarginal gyrus, BA 7, 40	1141	<0.001	4.30	51, -46, 44	R
	Precuneus, BA 40	339	0.010	4.21	-36, -46, 47	L
Within-memory	Middle & Superior FG, BA 9, 10	325	0.008	4.54	39, 41, 38	R
	Lingual gyrus, cuneus, calcarine	510	0.001	4.53	9, -91, 2	R
	Superior FG, dACC, pre-SMA	1066	<0.001	4.33	6, -4, 62	L/R
	Middle, Inferior FG, BA 9, 10, 45, 46	1003	<0.001	4.09	-42, 38, 2	L

856

857 **Table S5. Low-level visuomotor activity (Confidence/Follow ratings) in whole-brain search-**
858 **light classification accuracy map.** Significant activations at cluster-defining threshold $P<0.001$,
859 corrected for multiple comparisons at $P_{FWE}<0.05$
860

Label	**Voxels at *P*<0.001**	**P_{FWE} cluster-corrected**	**Peak *z*-score**	**Peak voxel MNI coordinates**	**Laterality**
Parietal lobe, occipital lobe, right insula & right middle FG	11240	<0.001	6.14	-39, -37, 47	L/R
Cerebellum	195	0.039	4.88	30, -70, -28	R

861

Figure S6. Confidence level-related activity in whole-brain searchlight classification accuracy map. Accuracy maps were masked to exclude low-level visuomotor regions (see Figure S8 & Table S5). Significant activations at cluster-defining threshold $P<0.001$, corrected for multiple comparisons at $P_{FWE}<0.05$

Classification	Label	Voxels at $P<0.001$	P_{FWE} cluster-corrected	Peak z-score	Peak voxel MNI coordinates	Laterality
Across-domain	Inferior FG, ACC, BA 11, 32	2434	<0.001	5.20	-21, 44, 2	L/R
	Superior and middle temporal gyri	176	<0.001	4.99	-57, -52, 20	L
	Middle temporal gyrus, anterior cerebellum	1193	<0.001	4.92	3, -46, -34	R
	Middle cingulate gyrus	46	0.017	4.85	-6, -10, 47	L/R
	Superior temporal gyrus	142	<0.001	4.68	66, -7, 5	R
	Precuneus, BA 7, 19	543	<0.001	4.68	-12, -61, 65	L/R
	Superior temporal gyrus	110	<0.001	4.68	-42, 26, -28	L
	Posterior cerebellum	96	<0.001	4.54	-9, -85, -22	L
	SMA, BA 6	84	0.001	4.53	-9, -10, 62	L/R
	Superior FG, dACC, pre-SMA	482	<0.001	4.39	-3, 11, 56	L/R
	Post-central gyrus	186	<0.001	4.33	-57, -19, 23	L
	Middle FG1	143	<0.001	4.28	36, 5, 26	R
	Fusiform and parahippocampal gyri	63	0.003	4.27	36, -19, -34	R
	Middle cingulate cortex	64	0.003	4.25	0, -31, 53	L/R
	Posterior cerebellum	53	0.008	4.10	-48, -58, -37	L
	Inferior FG	38	0.039	3.84	27, 23, -25	R
Within-perception	Superior FG, dACC/pre-SMA	185	0.032	3.88	-6, 29, 47	L/R
Within-memory	Inferior and middle FG, insula	503	<0.001	4.35	57, 23, 14	R
	dACC, pre-SMA	211	0.021	3.79	-9, 2, 47	L/R

868 **Figure S7. Metacognitive efficiency scores predictive of confidence level-related activity in**
869 **whole-brain searchlight classification accuracy maps.** Significant activations in the memory-
870 domain at cluster-defining threshold $P<0.001$, small-volume corrected for multiple comparisons
871 at $P_{FWE}<0.05$ using an anatomical precuneus mask. Significant activations in the perceptual do-
872 main at cluster-defining threshold $P<0.001$ (uncorrected).
873

Domain	**Label**	**Voxels at *P*<0.001**	**P_{FWE} small-volume cluster-corrected**	**Peak *z*-score**	**Peak voxel MNI coordinates**	**Laterality**
Memory	Precuneus, BA7	54	0.028	4.39	9, -70, 56	R

874

Domain	**Label**	**Voxels at *P*<0.001**	**P cluster-uncorrected**	**Peak *z*-score**	**Peak voxel MNI coordinates**	**Laterality**
Perception	Middle frontal gyrus, BA11	103	0.041	4.28	-15, 32, 11	L
	Cerebellum	111	0.035	4.16	-9, -76, -43	L/R
	Medial frontal gyrus, BA11	127	0.026	3.98	15, 50, -10	R

875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896

897 ***Supplementary References***
898

899 1. Fleming SM, Huijgen J, Dolan RJ. Prefrontal Contributions to Metacognition in Perceptual
900 Decision Making. J Neurosci. 2012 May 2;32(18):6117–25.
901 2. Fleck MS, Daselaar SM, Dobbins IG, Cabeza R. Role of Prefrontal and Anterior Cingulate
902 Regions in Decision-Making Processes Shared by Memory and Nonmemory Tasks. Cereb
903 Cortex. 2006;16(11):1623–30.
904 3. Hebart MN, Schriever Y, Donner TH, Haynes JD. The Relationship between Perceptual
905 Decision Variables and Confidence in the Human Brain. Cereb Cortex. 2014;26(1):118–30.
906 4. Baird B, Smallwood J, Gorgolewski KJ, Margulies DS. Medial and Lateral Networks in
907 Anterior Prefrontal Cortex Support Metacognitive Ability for Memory and Perception. J
908 Neurosci. 2013 Oct 16;33(42):16657–65.
909 5. Fleming SM, Weil RS, Nagy Z, Dolan RJ, Rees G. Relating Introspective Accuracy to Indi-
910 vidual Differences in Brain Structure. Science. 2010 Sep 16;329(5998):1541–3.
911 6. McCurdy LY, Maniscalco B, Metcalfe J, Liu KY, de Lange FP, Lau H. Anatomical Cou-
912 pling between Distinct Metacognitive Systems for Memory and Visual Perception. J Neuro-
913 sci. 2013 Jan 30;33(5):1897–906.
914 7. Andersson JLR, Hutton C, Ashburner J, Turner R, Friston K. Modeling Geometric Defor-
915 mations in EPI Time Series. Neuroimage. 2001 May;13(5):903–19.
916 8. Sladky R, Friston KJ, Tröstl J, Cunnington R, Moser E, Windischberger C. Slice-timing ef-
917 fects and their correction in functional MRI. Neuroimage. 2011 Sep 15;58(2):588–94.
918 9. Ashburner J, Friston KJ. Unified segmentation. Neuroimage. 2005 Jul;26(3):839–51.

919