Title Page

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

Abbreviated title: Domain-Specific Metacognition in Prefrontal Cortex

Authors: Jorge Morales[1*], Hakwan Lau[2,3,4], Stephen M. Fleming[5]

Affiliations:

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

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

Number of Figures: 7
Number of Tables: 5

Abstract: 170 words
Introduction: 624 words
Discussion: 973 words

Acknowledgements: We thank Aurelio Cortese for his help with the MVPA analyses. 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).

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 by human participants 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 predicted decoding accuracy of confidence activity 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.

SIGNIFICANCE STATEMENT: We use human neuroimaging to investigate the mechanisms for memory and perceptual metacognition. Metacognitive estimates must carry contextual information about the task if they are to be helpful in the control of behavior. Using univariate and decoding multivariate methods, we provide evidence that perceptual- and memory-specific metacognitive representations in right prefrontal cortex co-exist with generic confidence signals. Our findings reconcile previously conflicting results on the domain-specificity/generality of metacognition, and lay the groundwork for a mechanistic understanding of metacognitive judgments.

INTRODUCTION

Metacognition is the capacity to evaluate the success of one's cognitive processes in various domains, e.g. perception or memory (Flavell, 1979; Nelson and Narens, 1990; Metcalfe and Shimamura, 1994; Fleming et al., 2012a). Metacognitive ability can be assessed in the laboratory by quantifying the trial-by-trial correspondence between objective performance and subjective confidence (Galvin et al., 2003; Maniscalco and Lau, 2012; Overgaard and Sandberg, 2012; Fleming and Lau, 2014). Anatomical (Fleming et al., 2010; McCurdy et al., 2013; Maniscalco et al., 2017), functional (Fleck et al., 2006; Yokoyama et al., 2010; Fleming et al., 2012b; Baird et al., 2013; Hilgenstock et al., 2014) and neuropsychological (Flavell, 1979; Shimamura and Squire, 1986; Nelson and Narens, 1990; Metcalfe and Shimamura, 1994; Schnyer et al., 2004; Fleming et al., 2012a; 2014) 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 (Pouget et al., 2016; Fleming and Daw, 2017) 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 (de Gardelle and Mamassian, 2014). One solution to this problem is to maintain a global resource with access to arbitrary sensorimotor mappings (Holroyd et al., 2005; Heekeren et al., 2006; Cole et al., 2013). Candidate neural substrates for a domain-general resource are the frontoparietal and cingulo-opercular networks, known to be involved in arbitrary control operations (Cole et al., 2013). 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 (Gehring et al., 1993; Fleming et al., 2012b). 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 (Donoso et al., 2014), 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 (Galvin et al., 2003; Maniscalco and Lau, 2012; Fleming and Lau, 2014). Some studies have found that efficient metacognition in one task predicts good metacognition in another (McCurdy et al., 2013; Ais et al., 2016; Faivre et al., 2016), whereas others indicate the independence of metacognitive abilities (Kelemen et al., 2000; Baird et al., 2013; Vo et al., 2014). Recent studies employing bias-free measures of metacognition have identified differences in the neural correlates of memory and perceptual metacognition in both healthy subjects (Baird et al., 2013; McCurdy et al., 2013) and neuropsychological patients (Fleming et al., 2014). 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 on 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 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). None of the four paired t-tests (DOMAIN×STIMULUS) comparing performance between the Confidence and Follow conditions returned a significant difference (P>0.05). 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 the corresponding stimulus type in the memory task (shapes: PER M=73%, MEM M=67%; words: PER M=81%, MEM M=89%; Figure 1B). Critically, this en-

sured 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).

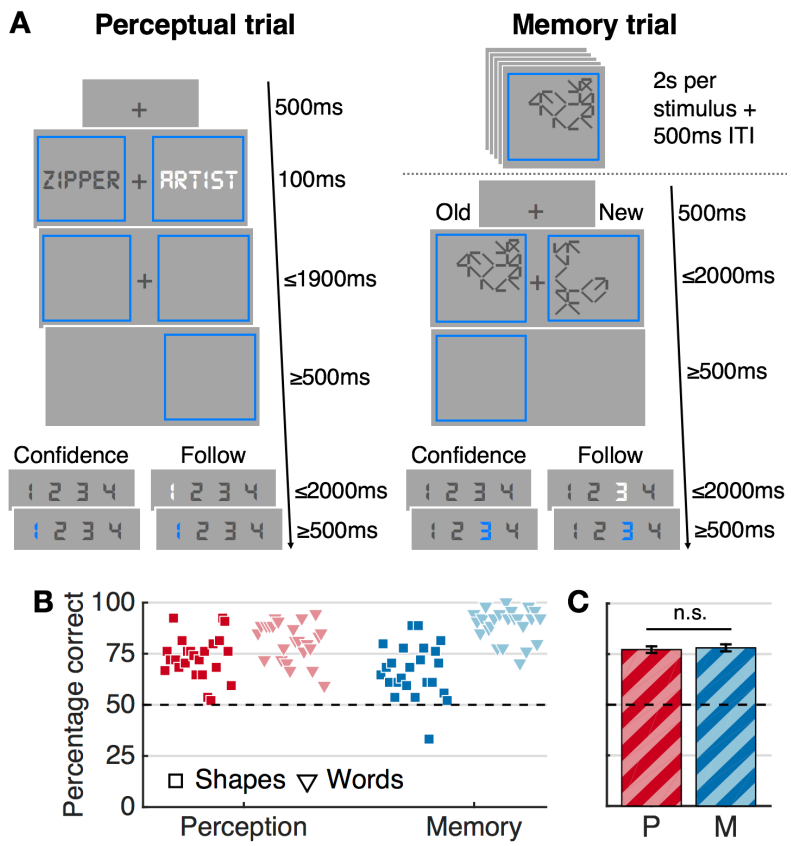

Figure 1. Task design & performance results. (A) Subjects performed two-alternative forced-choice discrimination tasks about perception and memory. In perception blocks, subjects selected the brighter of two stimuli. Memory blocks started with an encoding period and then subjects indicated in each trial 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)** 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.). n.s.= not significant; P=perception M=memory

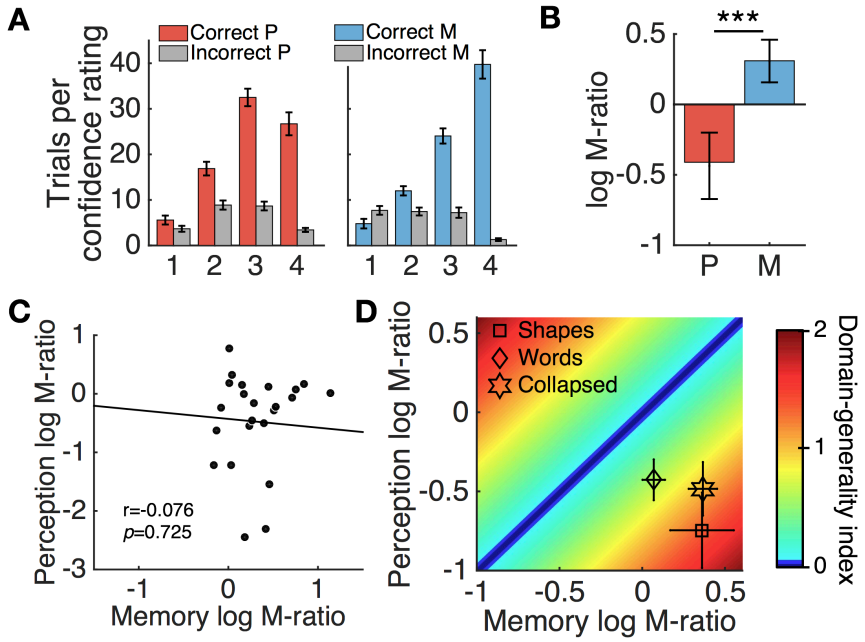

Figure 2. Metacognitive measures. (A) Mean number of correct and incorrect trials per confidence rating. **(B)** Metacognitive efficiency measured by log M-ratio (meta-*d*'/*d*'). Zero indicates that metacognitive sensitivity (meta-*d*') is equal to task sensitivity *d*' (i.e. the *d*' that would have 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)** Metacognitive efficiency scores obtained from single-subject Bayesian model fits were not correlated across 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. *** $P_{\theta>0}$ ~ 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(meta-d'/d'), a metacognitive efficiency measure derived from signal detection theory that assays the degree to which confidence ratings distinguish between correct and incorrect trials (Maniscalco and Lau, 2012; Fleming and Lau, 2014; Fleming, 2017). Metacognitive efficiency in the perceptual task was significantly lower than in the memory task ($P_{\theta>0}$ ~ 1; see Figure 2B & Methods for details), consistent with previous findings (Fleming et al., 2014). 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 scores in each domain for each participant (Fleming et al., 2014):

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

where M_P = perceptual meta-d'/d' and M_M = memory meta-d'/d'. 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 all analyses we focused on two features of metacognition-related activity: judgment-related (JR) activity (the difference between Confidence trials requiring a metacognitive judgment and the visuomotor control Follow condition) and confidence level-related (CLR) activity (the parametric relationship between confidence and activity).

Univariate results

Judgment-related activity. In standard univariate analyses, we found elevated activity in dACC/pre-SMA, bilateral insulae and superior and middle frontal gyri when contrasting the Confidence against the Follow condition, consistent with previous findings (Fleming et al., 2012b) (Figure 3A). Splitting the data by domain, a (Memory Confidence > Memory Follow) > (Perception Confidence > Perception Follow) interaction contrast revealed significant clusters of activity in middle cingulate gyrus, left insula, precuneus, left post-central gyrus and cerebellum (Figure 3B, blue). In a conjunction analysis, elevated activity for the Confidence > Follow condition was observed across both perception and memory trials in anterior cingulate and right insula (Figure 3B, green).

To further 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) (see Figure 3C & Methods), which previous studies have found to be recruited by perceptual and memory metacognition (Fleck et al., 2006; Fleming et al., 2010; 2012b; Baird et al., 2013; McCurdy et al., 2013). In a series of repeated measures 2×2 ANOVAS (condition × task) we found a main effect of greater activity on Confidence compared to Follow trials in all ROIs except precuneus, where instead we observed a main effect of task, with increased activity on memory trials (Figure 3D; condition main effect: dACC/pre-SMA, $F_{1,23}$=19.34, P=0.0002; left rlPFC, $F_{1,23}$=6.62, P=0.017; right rlPFC, $F_{1,23}$=9.28, P=0.006; PCUN: $F_{1,23}$=0.40, P=0.532; task main effect: dACC/pre-SMA, $F_{1,23}$=2.33, P=0.14; left rlPFC, $F_{1,23}$=0.95, P=0.34; right rlPFC, $F_{1,23}$=4.94, P=0.036; PCUN: $F_{1,23}$=36.78, P=3.47×10^{-6}). We found that the difference between Confidence and Follow trials was greater in memory than in perception trials in dACC/pre-SMA, recapitulating the whole-brain results (condition × task interaction; $F_{1,23}$=12.16, P=0.0019; paired t-test, MEM: T_{23}=5.47, P=0.0001, PER: T_{23}=1.92, P= 0.067). A similar interaction pattern was observed in precuneus (condition × task interaction: $F_{1,23}$=15.86, P=0.0006; paired t-test, MEM: T_{23}=2.43, P=0.023, PER: T_{23}=-1.54, P=0.136). There were no interactions in frontal regions (left rlPFC, $F_{1,23}$=0.07, P=0.795; right rlPFC, $F_{1,23}$=0.002, P=0.968). These results are compatible with previous findings indicating a distinctive contribution of

precuneus to memory metacognition (Baird et al., 2013; McCurdy et al., 2013).

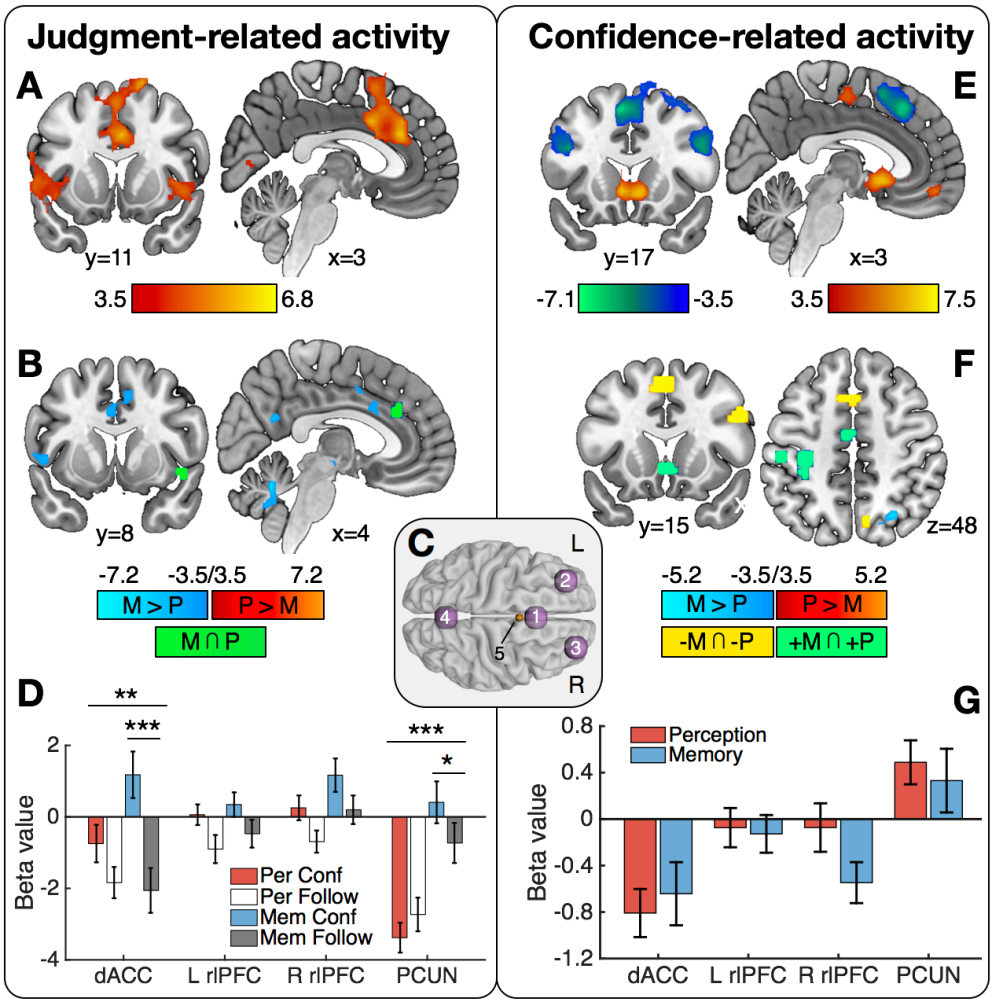

Figure 3. fMRI univariate analysis results. *Judgment-related activity*: **(A)** Whole-brain analysis of significant activation in the Confidence>Follow contrast; there were no significant clusters in the Follow>Confidence contrast. **(B)** (Memory Confidence > Memory Follow) > (Perception Confidence > Perception Follow) interaction contrast (blue). There were no significant clusters in the reverse contrast. The conjunction of Memory Confidence > Memory Follow and Perception Confidence > Perception Follow contrasts is indicated in green. See Table 1. **(C)** Spherical binary masks of four a priori regions of interest

[1=dACC, 2=L rlPFC, 3=R rlPFC, 4=PCUN] and one control region in the ventricles [5]. **(D)** Estimated mean beta values for judgment-related activity by domain in the four ROIs displayed in (C). *Confidence level-related activity*: **(E)** Whole-brain analysis of activity parametrically modulated by level of confidence. Hot colors indicate a positive correlation with confidence and cool colors a negative correlation. **(F)** Memory > Perception contrast (blue) testing for differences between the parametric effect of confidence by domain; there were no significant clusters in the Perception > Memory contrast. A conjunction analysis revealed shared activity that was positively (green) and negatively (yellow) correlated with confidence levels in both domains. See Table 2. **(G)** Estimated mean beta values of confidence level-related activity in the four ROIs displayed in (C). No significant difference in confidence-related activity between domains was found in any ROI (all $P>0.05$). In all whole-brain analyses activations are significant at a cluster-defining threshold $P<0.001$, uncorrected, corrected for multiple comparisons $P_{FWE}<0.05$; except for conjunction analyses in which we computed the intersection of two independent maps thresholded at $P<0.001$, uncorrected. Images displayed at $P<0.001$, uncorrected. Graded color bars reflect T-statistics. Error bars indicate s.e.m. *** $P<0.001$ ** $P<0.01$ *$P<0.05$. L=left; R=right; P=perception; M=memory.

Confidence level-related activity. We next sought to investigate the parametric relationship between confidence level and neural activity. We found activity in the posterior midline and ventromedial prefrontal cortex correlated positively with confidence (Figure 3E, hot colors). 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 (Fleck et al., 2006; Fleming et al., 2012b; Baird et al., 2013; Hebart et al., 2014) (Figure 3E, cool colors). When testing for differences between these parametric regressors by domain, a Memory > Perception contrast revealed a significant cluster of activity in right parietal cortex (Figure 3F, blue), while there was no significant activity in a Perception > Memory contrast. Shared positive correlations between confidence and activity in perception and memory trials were found in left pre- and post-central gyri, consistent with use of the right hand to provide confidence ratings (conjunction analysis; Figure 3F, green). Shared negative correlations with confidence were found in regions of right dorsolateral PFC and medial prefrontal cortex, overlapping with pre-SMA (Figure 3F, yellow).

Complementing the ROI analysis of JR-activity, we performed an ROI analysis of CLR-activity that recapitulated the whole-brain results.

We observed negative relationships between confidence and activity in dACC/pre-SMA and positive relationships in precuneus. Importantly, no significant differences in the parametric effect of confidence were found between domains in any of our a priori ROIs (Figure 3G; paired *t*-tests: dACC/pre-SMA, T_{23}=-0.47, P=0.643; left rlPFC, T_{23}=0.23, P=0.820; right rlPFC, T_{23}=1.62, P=0.119; PCUN: T_{23}=0.56, P=0.583). Together with the lack of domain-specific differences in confidence-related activity at the whole-brain level, such results are suggestive of an absence of domain-specificity in confidence-related activity. However, a lack of difference between univariate activation profiles is necessarily inconclusive. While metacognition-related activity may show overall levels of activation across tasks, distributed activity patterns in frontal and parietal areas may carry task-specific information (Hebart et al., 2014; Cole et al., 2016). We next turned to multivariate analysis methods to test this hypothesis.

Multivariate results

We performed a series of multi-voxel pattern analyses (MVPA) (Figure 5A) focused on both judgment-related activity and confidence level-related activity.

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 independent cross-validations for each domain and we tested for across-domain generalization using a cross-classification analysis (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; Figure 5B, dark colors). 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 5B). 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.

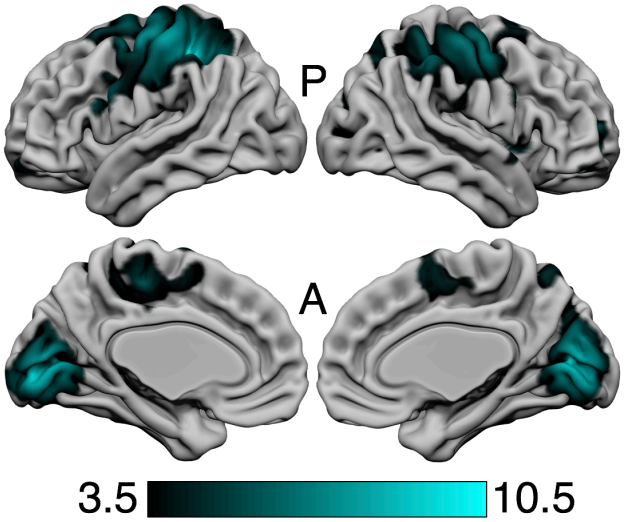

Figure 4. MVPA of low-level visuomotor features of the rating scale. Activations significant in a whole-brain searchlight analysis decoding ratings provided on Follow trials. This mask was used in whole-brain searchlight classification of confidence level-related activity (Figure 5E) to exclude visuomotor-related activations (see main text and Methods for details). Clusters are significant at P_{FWE}<0.05 cluster-level corrected for multiple comparisons using a cluster-defining threshold of P<0.001, uncorrected. Image displayed at P<0.001, uncorrected. The bar indicates T-statistics. P=posterior, A=anterior.

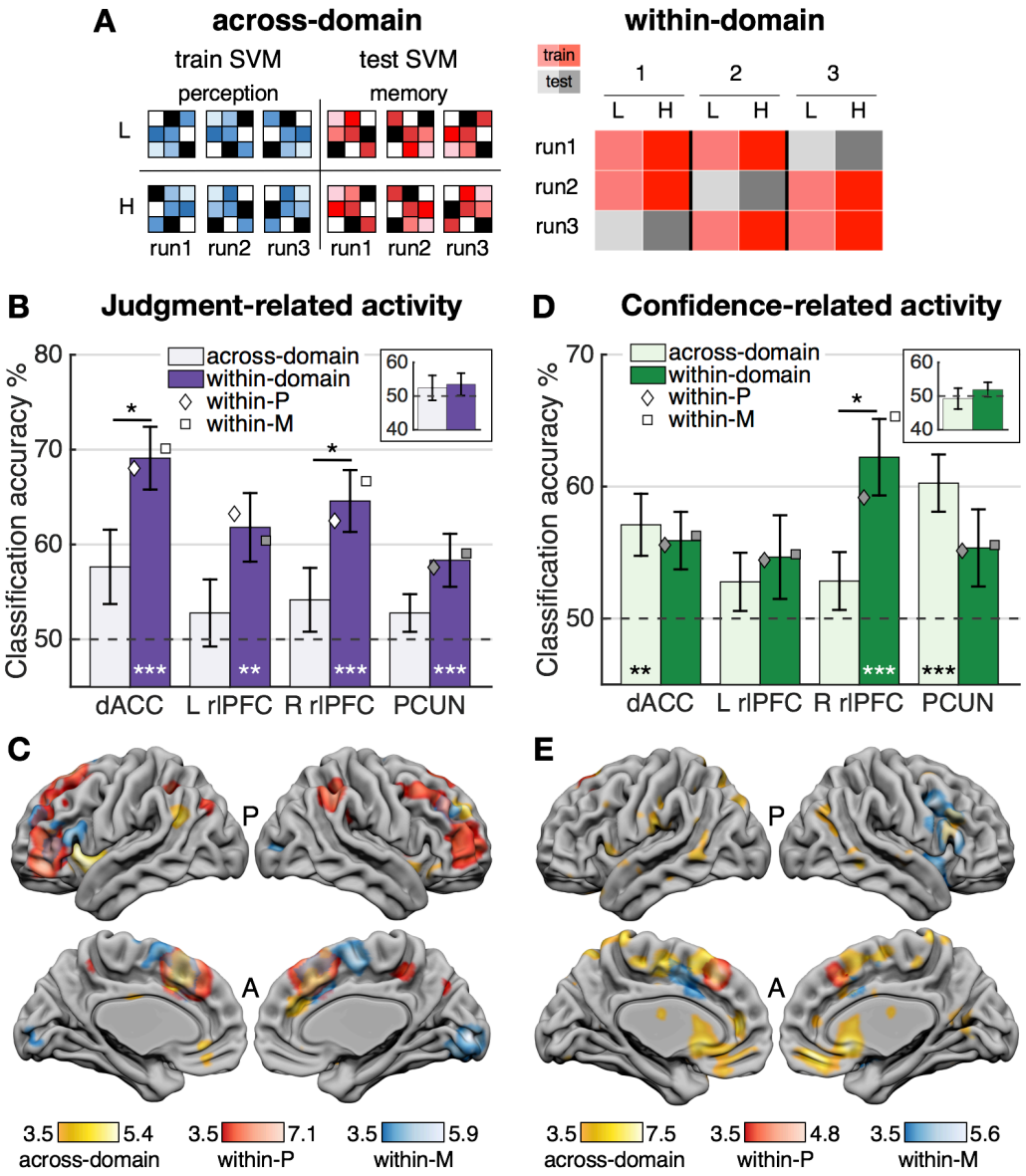

Figure 5. MVPA Results. *Design*: **(A)** *Left*: Across-domain classification design. Pattern vectors (run-wise beta images) from one domain were used to train an SVM decoder on two classes and then tested in a cross-classification of the same two classes using vectors from the other domain (and vice versa). Here we illustrate classification of low (L) and high (H) confidence levels. *Right:* Within-domain classification design. Pattern vectors of two classes (e.g. low and high confidence) pertaining to one domain were used to train a decoder in a leave-one-run-out design that was then tested in the left-out pair. The pro-

cess was iterated three times to test pairs from every run. An identical, independent cross-validation was performed on vectors from the other domain. *Judgment Related-Activity* **(B)** 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). **(C)** Searchlight analysis for same classifications as in (B) See Table 3. *Confidence Level Related-Activity* **(D)** Same as in (B) 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). **(E)** Searchlight analysis for the same classifications as in (D) exclusively masked for visuomotor-related activity. See Figure 4 and Table 4. Bars in (B) and (D) indicate means and error bars s.e.m. Dashed lines indicate chance classification (50%). Diamonds and squares indicate mean independent classification in perception and memory trials, respectively. White diamonds/squares indicate classification was significantly different from chance. All clusters in (C) and (E) 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. P=perception; M=memory; A=anterior; P=posterior. *** P≤0.001 ** P≤0.01 * P<0.05; one-sample t-tests are Bonferroni corrected.

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 (Hebart et al., 2015). 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 5C). 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 3A).

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, ROI univariate analyses did not reveal any differences in confidence-related activity between domains (Figure 3G). 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, significant within-domain classification 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 5D, 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). The mean classification accuracy in this ROI was 62% (PER=59%, MEM=65%), notably above a recently estimated median 55% for decoding task-relevant information in frontal regions (Bhandari et al., 2017). Importantly, classification accuracy in this ROI also differed from the corresponding across-domain classification accuracy (paired *t*-test T_{21}=2.37, P=0.027; Figure 5D, dark colors). Together, these results suggest the co-existence 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- but not across-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 4). 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 5E). We found widespread across-domain classification of confidence (yellow) in a predominantly midline network including a large cluster encompassing dACC/pre-SMA, ventromedial PFC (vmPFC), and striatum. 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 to objective accuracy. To further address the question of how confidence judgments may relate to activity patterns, we examined the relationship between objective task accuracy and confidence. Previous work suggests that the neural basis (and associated activation patterns) related to confidence and performance may be partly distinct (Rounis et al., 2010; Cortese et al., 2016). Specifically, we tested for the ability of CLR-activity patterns to predict objective performance (correct/incorrect) on Follow trials and vice versa in a cross-classification analysis (collapsed across domain). 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 predict objective accuracy levels in Follow trials above chance (Figure 6; light blue). We then compared these decoding scores to a leave-one-run-out cross-validation decoding analysis of low vs. high confidence on Confidence trials only (collapsed by domain; Figure 6, dark blue). Consistent with the domain-general analyses reported in Figure 5D (light colors), this decoder was unable to classify confidence patterns of activity in the frontal regions (left P=0.12; right P=0.40), but was significantly above chance in dACC/pre-SMA (T_{21}= 2.58, P=0.003) and precuneus (T_{21}= 2.48, P=0.002). Critically, in precuneus, confidence classification was significantly greater than confidence-accuracy generalization, which was at chance (paired t-test T_{21}= 3.44, P=0.002). This result indicates that confidence-related patterns in precuneus are unable to predict objective accuracy, consistent with a partly distinct coding of information relevant to task performance and confidence. In contrast, in dACC/pre-SMA, confidence and accuracy could be predicted from common patterns of activation.

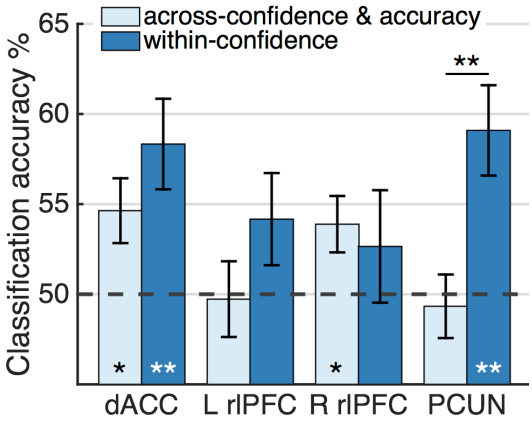

Figure 6. Generalization of confidence-related activity to objective accuracy. Light blue bars denote mean cross-classification accuracy results obtained from training a decoder on CLR-activity and predicting objective accuracy (correct / incorrect) in the Follow condition (and vice versa). Dark blue bars denote decoding accuracy for a leave-one-out cross-validation of low and high confidence on Confidence trials only. Bars indicate group means and bars s.e.m. Dotted line indicates chance level. * $P<0.05$ ** $P<0.01$

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 classification 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 7). 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 (Fleming et al., 2010), anatomical (McCurdy et al., 2013) and lesion studies (Fleming et al., 2014) that have linked individual variation in metacognitive accuracy to precuneus and PFC in memory and perception, respectively.

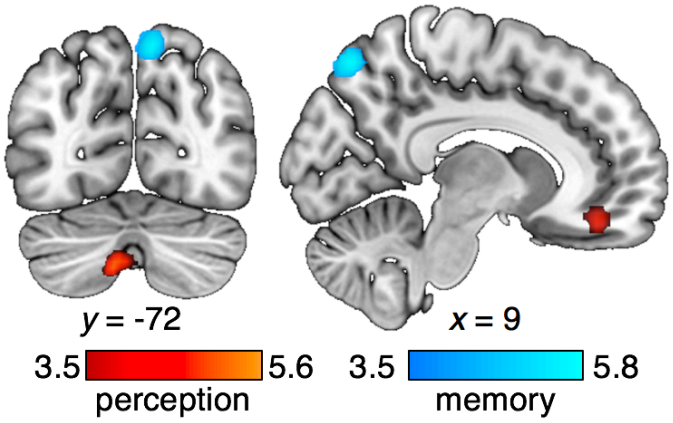

Figure 7. Relationship between metacognitive efficiency and confidence classification. Individual metacognitive efficiency scores predicted within-domain classification accuracy of confidence level-related activity. In memory trials, they predicted classification accuracy in a cluster in right precuneus (blue). In perceptual trials, individual metacognitive efficiency scores predicted confidence classification in bilateral ventromedial prefrontal cortex and cerebellum (red). Images displayed at $P<0.001$, uncorrected. See main text for statistical analyses. Color bars indicate T-statistic. See Table 5.

DISCUSSION

When performing a cognitive task, confidence estimates allow for comparisons of performance across a range of different scenarios (de Gardelle and Mamassian, 2014). 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 (McCurdy et al., 2013), 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 and

objective accuracy across memory and perceptual tasks. Previous studies of metacognition have implicated the cingulo-opercular network in tracking confidence level (Fleck et al., 2006; Fleming et al., 2012b; De Martino et al., 2013; Hebart et al., 2014; Hilgenstock et al., 2014). 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 (Scheffers and Coles, 2000; Boldt and Yeung, 2015). The activity in this region was also consistently elevated by the requirement for a metacognitive judgment (Fleming et al., 2012b), although the pattern of these increases did not generalize across tasks. Notably, while both dACC/pre-SMA and precuneus both showed significant domain-general decoding of confidence, in precuneus these patterns did not generalize to also predict changes in objective accuracy, consistent with a partial distinction between neural substrates of confidence and first-order performance revealed through multi-voxel neurofeedback (Cortese et al., 2016).

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 (Figure 5D): 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 (Donoso et al., 2014; Purcell and Kiani, 2016). 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 (Purcell and Kiani, 2016). 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 me-

dial and lateral PFC in the hierarchy of control (Kerns et al., 2004; Tang et al., 2016). 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 3D). Individual metacognitive efficiency scores in memory trials predicted classification accuracy in precuneus (Figure 7), while individual differences in metacognitive efficiency scores in perceptual trials predicted classification accuracy in ventromedial prefrontal cortex (albeit the latter at an uncorrected threshold). These findings are consistent with the medial parietal cortex making a disproportional contribution to memory metacognition (Baird et al., 2013; McCurdy et al., 2013), and offer a potential explanation for a decrease in perceptual, but not memory, metacognitive efficiency seen in patients with frontal lesions (Fleming et al., 2012b). 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 (Fleming et al., 2010; McCurdy et al., 2013).

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 (Macpherson, 2011). 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 (Faivre et al., 2016). However, it is an open question whether more fine-grained 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 (Poldrack et al., 2017). The study protocols were approved by the local Institutional Review Board.

Task

Experiments were programmed in MATLAB (MathWorks) using Psychtoolbox (Brainard, 1997). 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 (Fleming et al., 2010; 2012b; 2014). 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 (García-Pérez, 1998) 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 centrality 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 (Wilson, 1988). 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 (Nelson et al., 2004).

Behavioral data analysis

Data analysis was performed in MATLAB and statistical analysis in RStudio (R Studio Team, 2015). 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') (Maniscalco and Lau, 2012; Fleming and Lau, 2014). We used hierarchical Bayesian estimation to incorporate subject-level uncertainty in group-level parameter estimates (Fleming, 2017). Certainty on this parameter was determined by computing the 95% high-density interval (HDI) from the posterior samples (Kruschke, 2010). For correlation and individual differences analyses we used single-subject Bayesian model fits. 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.

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 (Andersson et al., 2001) and then slice-timing corrected (Sladky et al., 2011). 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 (Ashburner and Friston, 2005). 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 judgment-related (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 confidence level-related (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. For conjunction analyses of activations common to both domains, second-level maps thresholded at $P<0.001$ uncorrected were intersected to reveal regions of shared statistically significant JR- and CLR-activity. 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 (ROI), 12mm spheres were centered at MNI coordinates identified from previous literature (Figure 3C). 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 (Fleming et al.,

2012b). The mask for precuneus (PCUN) [0, -64, 24] was based on (McCurdy et al., 2013). 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]. Beta values were extracted from subjects' Confidence>Follow contrast and confidence parametric modulator contrast images for the JR and CLR univariate analyses, respectively. For the multi-voxel pattern analyses, the ROI spheres were used as masks during decoding.

Multi-voxel pattern analysis

Multi-voxel pattern analysis (MVPAs) were carried out in MATLAB using The Decoding Toolbox (Hebart et al., 2015). We classified run-wise beta images from general linear models (GLMs) modeling JR- and CLR-activity in ROI and whole-brain searchlight analyses. ROI MVPAs were performed on normalized, smoothed images. Previous work has shown that these preprocessing steps have minimal impact on support vector machine (SVM) classification accuracy, while allowing meaningful comparison across subject-specific differences in anatomy, as in standard fMRI analyses (Kamitani and Sawahata, 2010; Op de Beeck, 2010). A single accuracy value per subject, per condition was extracted and used for group analysis.

Searchlight analyses used 12mm-radius spheres centered around a given voxel, for all voxels, on spatially realigned and slice-timing corrected images from each subject to create whole-brain accuracy maps. For group-level analyses, these individual searchlight maps were spatially normalized and smoothed using a Gaussian kernel (8mm FWHM) and entered into one-sample *t*-tests against chance accuracy (Hebart et al., 2014; 2015). Whole-brain cluster inference was carried out in the same manner as in univariate analysis.

Prior to decoding, for JR-activity classification we modeled the rating periods in the Confidence and Follow conditions in two regressors of interest per run. For CLR-activity, we collapsed ratings 1 and 2 into a low-confidence regressor and ratings 3 and 4 into a high-confidence regressor to allow binary classification. Despite being encouraged to use the four levels of confidence in every run, two subjects did not provide ratings for one of the classes in at least one run and were left out from the analysis to avoid entering unbalanced training data into the classifier. The main result of the ROI analysis (Figure 5D) is not affected when the unbalanced data of

these two subjects are included: within-domain classification accuracy in right rlPFC (M=62.22%) is significantly above chance (T_{23}= 4.22; P=0.0003) and different from across-domain classification accuracy (M=52.84%; paired t-test: T_{23}= 2.54; P=0.018). The remaining parameters were specified as in the univariate case.

In independent across-domain classifications, we used the run-wise beta images reflecting JR- and CLR-activity as pattern vectors in a linear support vector classification model (as implemented in LIBSVM). We assigned each vector from each domain a label corresponding to the classes Confidence (1) and Follow (-1) in the JR-analysis and Low Confidence (-1) and High Confidence (1) in CLR-analysis. We trained an SVM with the vectors from one domain (3 per class, 6 in total) and tested the decoder on the 6 vectors from the other domain (and vice versa) (Figure 5A; left). We report the average classification accuracies of these two-way cross-classifications.

For within-domain classifications, we ran independent leave-one-run-out cross-validations for each domain on JR-activity (Confidence vs Follow) and CLR-activity (Low vs High confidence). The pattern vectors from two of the three runs in each domain were used to train an SVM to predict the same classes in the vectors from the left-out run. We compared the true labels of the left-out run with the labels predicted by the model and iterated this process for the other two runs to calculate a mean cross-validated accuracy independently for each domain (Figure 5A; right).

To test for the ability of confidence-related activity patterns to predict objective performance in the absence of confidence reports, we performed a cross-classification analysis in which a decoder trained on Confidence trials (Low vs High Confidence) was tested on pattern vectors from Follow trials (Incorrect vs Correct), and vice versa (collapsed across domain). This confidence-accuracy generalization score was compared to a leave-one-run-out cross-validation analysis decoding Low vs. High confidence on Confidence trials only (collapsed across domain). Together, these scores characterize whether a particular set of patterns are specific to confidence, or also generalize to predict objective performance (Cortese et al., 2016).

Individual differences

Metacognitive efficiency scores (log meta-d'/d') were estimated independently for the perceptual and memory tasks, together with a single score collapsed across domains. These scores were inserted as covariates in second-level analyses of within-perception, within-memory and across-domain classifications of confidence level-related activity, respectively. A contrast was applied to estimate the parametric relationship of metacognitive efficiency and decoding success for each subject. An anatomical map of the precuneus (Harvard-Oxford probabilistic atlas) was used for small-volume correction of within-memory decoding results.

REFERENCES

Ais J, Zylberberg A, Barttfeld P, Sigman M (2016) Individual consistency in the accuracy and distribution of confidence judgments. Cognition 146:377–386.

Andersson JLR, Hutton C, Ashburner J, Turner R, Friston K (2001) Modeling Geometric Deformations in EPI Time Series. Neuroimage 13:903–919.

Ashburner J, Friston KJ (2005) Unified segmentation. Neuroimage 26:839–851.

Baird B, Smallwood J, Gorgolewski KJ, Margulies DS (2013) Medial and Lateral Networks in Anterior Prefrontal Cortex Support Metacognitive Ability for Memory and Perception. J Neurosci 33:16657–16665.

Bhandari A, Gagne C, Badre D (2017) Just above chance: is it harder to decode information from human prefrontal cortex BOLD signals? bioRxiv:1–28.

Boldt A, Yeung N (2015) Shared Neural Markers of Decision Confidence and Error Detection. J Neurosci 35:3478–3484.

Brainard DH (1997) The Psychophysics Toolbox. Spat Vis 10:433–436.

Cole MW, Ito T, Braver TS (2016) The Behavioral Relevance of Task Information in Human Prefrontal Cortex. Cereb Cortex 26:2497–2505.

Cole MW, Reynolds JR, Power JD, Repovs G, Anticevic A, Braver TS (2013) Multi-task connectivity reveals flexible hubs for adaptive task control. Nature Neurosci 16:1348–1355.

Cortese A, Amano K, Koizumi A, Kawato M, Lau H (2016) Multivoxel neurofeedback selectively modulates confidence without changing perceptual performance. Nat Commun 7.

de Gardelle V, Mamassian P (2014) Does Confidence Use a Common Currency Across Two Visual Tasks? Psychol Sci.

De Martino B, Fleming SM, Garrett N, Dolan RJ (2013) Confidence in val-

ue-based choice. Nature Neurosci 16:105–110.
Donoso M, Collins AGE, Koechlin E (2014) Foundations of human reasoning in the prefrontal cortex. Science 344:1481–1486.
Faivre N, Filevich E, Solovey G, Kuhn S, Blanke O (2016) Behavioural, modeling, and electrophysiological evidence for domain-generality in human metacognition. bioRxiv.
Flavell JH (1979) Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. American Psychologist 34:906–911.
Fleck MS, Daselaar SM, Dobbins IG, Cabeza R (2006) Role of Prefrontal and Anterior Cingulate Regions in Decision-Making Processes Shared by Memory and Nonmemory Tasks. Cereb Cortex 16:1623–1630.
Fleming SM (2017) Hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. Neurosci Consc 3.
Fleming SM, Daw ND (2017) Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. Psychol Rev 124:91–114.
Fleming SM, Dolan RJ, Frith CD (2012a) Metacognition: computation, biology and function. Phil Trans R Soc B 367:1280–1286.
Fleming SM, Huijgen J, Dolan RJ (2012b) Prefrontal Contributions to Metacognition in Perceptual Decision Making. J Neurosci 32:6117–6125.
Fleming SM, Lau H (2014) How to measure metacognition. Front Hum Neurosci 8.
Fleming SM, Ryu J, Golfinos JG, Blackmon KE (2014) Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. Brain 137:2811–2822.
Fleming SM, Weil RS, Nagy Z, Dolan RJ, Rees G (2010) Relating Introspective Accuracy to Individual Differences in Brain Structure. Science 329:1541–1543.
Galvin SJ, Podd JV, Drga V, Whitmore J (2003) Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. Psychon Bull Rev 10:843–876.
García-Pérez MA (1998) Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties. Vision Res 38:1861–1881.
Gehring WJ, Goss B, Coles MGH, Meyer DE, Donchin E (1993) A Neural System for Error Detection and Compensation. Psychol Sci 4:385–390.
Hebart MN, Görgen K, Haynes JD (2015) The Decoding Toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. Front Neuroinform 8.
Hebart MN, Schriever Y, Donner TH, Haynes JD (2014) The Relationship between Perceptual Decision Variables and Confidence in the Human Brain. Cereb Cortex 26:118–130.
Heekeren HR, Marrett S, Ruff DA, Bandettini PA, Ungerleider LG (2006)

Involvement of human left dorsolateral prefrontal cortex in perceptual decision making is independent of response modality. Proc Natl Acad Sci USA 103:10023–10028.
Hilgenstock R, Weiss T, Witte OW (2014) You'd Better Think Twice: Post-Decision Perceptual Confidence. Neuroimage 99:323–331.
Holroyd CB, Yeung N, Coles MGH, Cohen JD (2005) A Mechanism for Error Detection in Speeded Response Time Tasks. J Exp Psychol Gen 134:163–191.
Kamitani Y, Sawahata Y (2010) Spatial smoothing hurts localization but not information: Pitfalls for brain mappers. Neuroimage 49:1949–1952.
Kelemen WL, Frost PJ, Weaver CA (2000) Individual differences in metacognition: evidence against a general metacognitive ability. Mem Cognit 28:92–107.
Kerns JG, Cohen JD, Stenger VA, Carter CS (2004) Prefrontal cortex guides context-appropriate responding during language production. Neuron 43:283–291.
Kruschke J (2010) Doing Bayesian Data Analysis. Academic Press.
Macpherson F (2011) Introduction: Individuating the Senses. In: The Senses (Macpherson F, ed), pp 3–42 The Senses: Classical and Contemporary Readings. Oxford: Oxford University Press.
Maniscalco B, Lau H (2012) A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. Conscious Cogn 21:422–430.
Maniscalco B, McCurdy LY, Odegaard B, Lau H (2017) Limited Cognitive Resources Explain a Trade-Off between Perceptual and Metacognitive Vigilance. J Neurosci 37:1213–1224.
McCurdy LY, Maniscalco B, Metcalfe J, Liu KY, de Lange FP, Lau H (2013) Anatomical Coupling between Distinct Metacognitive Systems for Memory and Visual Perception. J Neurosci 33:1897–1906.
Metcalfe J, Shimamura AP eds. (1994) Metacognition. Cambridge, MA: MIT Press.
Nelson DL, McEvoy CL, Schreiber TA (2004) The University of South Florida free association, rhyme, and word fragment norms. Behav Res Methods Instrum Comput 36:402–407.
Nelson TO, Narens L (1990) Metamemory: A Theoretical Framework and New Findings. In: The Psychology of Learning and Motivation (Bower G, ed), pp 125–173. New York: Academic Press.
Op de Beeck HP (2010) Against hyperacuity in brain reading: spatial smoothing does not hurt multivariate fMRI analyses? Neuroimage 49:1943–1948.
Overgaard M, Sandberg K (2012) Kinds of access: different methods for report reveal different kinds of metacognitive access. Phil Trans R Soc B

367:1287–1296.
Poldrack RA, Baker CI, Durnez J, Gorgolewski KJ, Matthews PM, Munafò MR, Nichols TE, Poline J-B, Vul E, Yarkoni T (2017) Scanning the horizon: towards transparent and reproducible neuroimaging research. Nat Rev Neurosci 18:115–126.
Pouget A, Drugowitsch J, Kepecs A (2016) Confidence and certainty: distinct probabilistic quantities for different goals. Nature Neurosci 19:366–374.
Purcell BA, Kiani R (2016) Hierarchical decision processes that operate over distinct timescales underlie choice and changes in strategy. Proc Natl Acad Sci USA 113:E4531–E4540.
R Studio Team (2015) RStudio: Integrated Development for R. Available at: http://www.rstudio.com.
Rounis E, Maniscalco B, Rothwell JC, Passingham RE, Lau H (2010) Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. Cogn Neurosci 1:165–175.
Scheffers MK, Coles M (2000) Performance monitoring in a confusing world: error-related brain activity, judgments of response accuracy, and types of errors. J Exp Psychol Human 26:141–151.
Schnyer DM, Verfaellie M, Alexander MP, LaFleche G, Nicholls L, Kaszniak AW (2004) A role for right medial prefrontal cortex in accurate feeling-of-knowing judgments: evidence from patients with lesions to frontal cortex. Neuropsychologia 42:957–966.
Shimamura AP, Squire LR (1986) Memory and metamemory: a study of the feeling-of-knowing phenomenon in amnesic patients. J Exp Psychol Learn 12:452–460.
Sladky R, Friston KJ, Tröstl J, Cunnington R, Moser E, Windischberger C (2011) Slice-timing effects and their correction in functional MRI. Neuroimage 58:588–594.
Tang H, Yu H-Y, Chou C-C, Crone NE, Madsen JR, Anderson WS, Kreiman G (2016) Cascade of neural processing orchestrates cognitive control in human frontal cortex. eLife 5.
Vo VA, Li R, Kornell N, Pouget A, Cantlon JF (2014) Young Children Bet on Their Numerical Skills: Metacognition in the Numerical Domain. Psychol Sci 25:1712–1721.
Wilson M (1988) MRC Psycholinguistic Database: Machine-usable dictionary, version 2.00. Behav Res Methods Instrum Comput 20:6–10.
Yokoyama O, Miura N, Watanabe J, Takemoto A, Uchida S, Sugiura M, Horie K, Sato S, Kawashima R, Nakamura K (2010) Right frontopolar cortex activity correlates with reliability of retrospective rating of confidence in short-term recognition memory performance. Neurosci Res 68:199–206.

TABLES

Table 1. Univariate fMRI analysis - judgment-related activity interacted with domain. Significant activations at cluster-defining threshold $P<0.001$, uncorrected, corrected for multiple comparisons at $P_{FWE}<0.05$. Conjunction of significant activations at cluster-defining threshold $P<0.001$, uncorrected, of Memory (C>F) and Perception (C>F) contrasts. See Figure 3B. C=confidence, F=follow

Contrast	Label	Voxels at $P<0.001$	P_{FWE} cluster-corrected	Peak z-score	Peak voxel MNI coordinates	Laterality
Memory (C > F) > Perception (C > F)	Cerebellum	70	0.015	5.17	3, -58, -25	L/R
	Insula	109	0.001	4.89	-33, -10, -7	L
	Posterior cingulate, precuneus	84	0.006	4.29	3, -58, 23	L/R
	Post-central gyrus, BA3	99	0.002	4.25	-45, -31, 62	L
	Hippocampus, parahippocampal gyrus, fusiform area	100	0.002	4.13	-21, -37, -16	L
	Thalamus	66	0.019	4.10	9, -25, -7	R
	Middle and anterior cingulate gyrus, SMA, BA24 & 32	194	<0.001	4.09	-6, 5, 32	L/R
	Inferior frontal gyrus, BA47	61	0.026	4.05	-42, 20, -4	L
Conjunction	Cingulate gyrus, BA32	12			6, 32, 29	R
	Insula	7			45, 11, -7	R

Table 2. Univariate fMRI analysis - confidence level-related activity interacted with domain. Significant activations at cluster-defining threshold $P<0.001$, uncorrected, corrected for multiple comparisons at $P_{FWE}<0.05$. Conjunction of significant activations at cluster-defining threshold $P<0.001$, uncorrected, of Memory (M) and Perception (P) contrasts of positive and negative correlations with confidence level. See Figure 3F.

Contrast	Label	Voxels at $P<0.001$	P_{FWE} cluster-corrected	Peak z-score	Peak voxel MNI coordinates	Laterality
Memory > Perception	Precuneus, BA7	93	0.003	4.21	33, -70, 20	R
Conjunction (+M ∩ +P)	Precentral & postcentral gyri, BA6, 4, 3	167			-30, -25, 44	L
	Post-central gyrus	27			-27, -46, 59	L
	SMA, BA6	21			-3, -10, 50	L/R
	Anterior cingulate, BA25	16			-30, -25, 44	L/R
	Cerebellum	7			9, -58, -13	R
	Post-central gyrus	5			-48, -22, 44	R
Conjunction (-M ∩ -P)	Middle frontal gyrus	29			45, 26, 20	R
	Pre-SMA, BA8	19			0, 14, 50	L/R

Table 3. Judgment-related activity obtained from whole-brain searchlight classification analyses. Significant activations at cluster-defining threshold $P<0.001$, uncorrected, corrected for multiple comparisons at $P_{FWE}<0.05$. See Figure 5C.

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

Table 4. Confidence level-related activity obtained from whole-brain searchlight classification analyses. Accuracy maps were masked to exclude visuomotor-related activity (see Figure 4). Significant activations at cluster-defining threshold $P<0.001$, uncorrected, corrected for multiple comparisons at $P_{FWE}<0.05$. See Figure 5E.

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

Table 5. Regions of confidence level-related activity predictive of metacognitive efficiency across individuals. Significant activations in the memory domain were observed at a cluster-defining threshold $P<0.001$, uncorrected, small-volume corrected for multiple comparisons at $P_{FWE}<0.05$ using an anatomical precuneus mask. Activations in the perceptual domain were observed at an uncorrected threshold of $P<0.001$. See Figure 7.

Domain	Label	Voxels at $P<0.001$	P value	Peak z-score	Peak voxel MNI coordinates	Laterality
Memory	Precuneus, BA7	54	0.028 (P_{FWE} small-volume cluster-corrected)	4.39	9, -70, 56	R
Perception	Middle frontal gyrus, BA11	103	0.041 (uncorrected)	4.28	-15, 32, 11	L
	Cerebellum	111	0.035 (uncorrected)	4.16	-9, -76, -43	L/R
	Medial frontal gyrus, BA11	127	0.026 (uncorrected)	3.98	15, 50, -10	R