Title: Domain-General and Domain-Specific Patterns of Activity Supporting 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[5,6]

Affiliations:

[1] Department of Philosophy, Columbia University, New York, NY, USA
[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 Centre for Human Neuroimaging, University College London, London, UK
[6] Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London, London, UK

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

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). The Wellcome Centre for Human Neuroimaging is supported by core funding from the Wellcome Trust (Grant No. 203147/Z/16/Z).

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 anterior prefrontal 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. 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 processes supporting memory and perceptual metacognition. It remains controversial whether metacognition relies on a global resource that is applied to different tasks, or whether self-evaluative processes are specific to particular tasks. Using multivariate decoding methods, we provide evidence that perceptual- and memory-specific metacognitive representations 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 (Shimamura and Squire, 1986; Schnyer et al., 2004; Fleming et al., 2014) evidence indicates specific neural substrates (especially in frontolateral, frontomedial, and parietal regions) contribute to metacognition across a range of task domains, including perception and memory. 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,

46 known to be involved in arbitrary control operations (Cole et al., 2013). In particular, the
47 dorsomedial prefrontal cortex (encompassing the paracingulate cortex and pre-supplementary
48 motor area) has been implicated in representing confidence, monitoring conflict, and detecting
49 errors across a range of tasks (Gehring et al., 1993; Botvinick et al., 2004; Ridderinkhof et al.,
50 2004; Fleming et al., 2012b). On the other hand, if the system only had access to generic
51 confidence signals, appropriate switching between particular tasks or strategies on the basis of
52 their expected success would be compromised. Functional imaging evidence implicates human
53 anterior prefrontal cortex in tracking the reliability of specific alternative strategies during
54 decision-making (Donoso et al., 2014), and such regions may also support domain-specific
55 representations of confidence.

56 Current behavioral evidence of a shared resource for metacognition is ambiguous, in part
57 due to the difficulty of distilling metacognitive processes from those supporting primary task
58 performance (Galvin et al., 2003; Maniscalco and Lau, 2012; Fleming and Lau, 2014). Some
59 studies have found that efficient metacognition in one task predicts good metacognition in
60 another (McCurdy et al., 2013; Ais et al., 2016; Faivre et al., 2017; Ruby et al., 2017; Samaha
61 and Postle, 2017), whereas others indicate the independence of metacognitive abilities (Kelemen
62 et al., 2000; Baird et al., 2013; Vo et al., 2014). Recent studies employing bias-free measures of
63 metacognition have identified differences in the neural correlates of memory and perceptual
64 metacognition in both healthy subjects (Baird et al., 2013; McCurdy et al., 2013) and
65 neuropsychological patients (Fleming et al., 2014). However, the study of behavioral individual
66 differences provides only indirect evidence of the neural and computational architecture
67 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 tracks confidence level. Together, our findings reveal the co-existence of generic and specific confidence representations, consistent with a computational hierarchy underpinning effective metacognition.

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 study protocols were approved by the local Institutional Review Board. 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). Due to behavioral and in-scanner motion cut-off criteria, six subjects were excluded from further analysis (details below). We present the results of twenty-four subjects whose data were fully analyzed.

91 ***Experimental and Task Design***

92 The experiment had a 2×2×2 design: CONDITION (CONFIDENCE/FOLLOW) × TASK DOMAIN
93 (PERCEPTION/MEMORY) × STIMULUS TYPE (SHAPES/WORDS). It consisted of six scanner runs, each
94 with eight 9-trial mini-blocks (72 trials per run, 432 trials in total). Perceptual and memory two-
95 alternative forced-choice (2-AFC) tasks were presented in separate, interleaved runs (three runs
96 per task; order counterbalanced across subjects). In each run, there were four pairs of mini-
97 blocks from the Confidence and Follow conditions. To avoid stimulus confounds, two different
98 types of stimulus were used throughout the experiment. In each run, two pairs of
99 Confidence/Follow mini-blocks used words and the remaining two pairs used abstract shapes
100 (interleaved and order counterbalanced across runs).

101 In the perceptual task, subjects were asked to report the brighter of two stimuli on each
102 trial. In the memory task, subjects began each mini-block by learning a set of nine consecutively-
103 presented stimuli. A stimulus from this set was then presented on each subsequent trial (in
104 randomized order) alongside a new stimulus. Subjects' task was to identify the studied stimulus.
105 In mini-blocks from the Confidence condition, subjects had to rate their confidence in their
106 performance in each trial by selecting a number from a 1-to-4 scale. In mini-blocks from the
107 Follow condition, subjects had to "follow the computer" in each trial by pressing the button
108 corresponding to the highlighted number irrespective of their confidence. The highlighted
109 number was yoked to their ratings in the previous Confidence mini-block (randomized
110 presentation order) to ensure similar low-level visuomotor characteristics in both conditions for
111 any given pair of mini-blocks.

112 Subjects were reminded at the beginning of each mini-block of the condition, task, and
113 stimulus type that would follow. They used two fingers of their right hand to respond on an MRI-

114 compatible button box: left stimulus (index) and right stimulus (middle). For confidence ratings,
115 they used four fingers: 1 (index), 2 (middle), 3 (ring) and 4 (little). If subjects failed to provide
116 either type of response within the allotted time (see Figure 1A for details), the trial was missed
117 and an exclamation mark was displayed for the remainder of the trial. Failing to press the
118 highlighted number counted as a missed trial.

119 Prior to entering the scanner, participants were familiarized with the tasks and the
120 confidence rating scale. After computing independent brightness thresholding for words and
121 abstract shapes, subjects practiced one of each mini-block type (i.e. 8 mini-blocks). Instructions
122 emphasized that confidence ratings should reflect relative confidence and participants were
123 encouraged to use all ratings. The whole experiment lasted ~1.5 hours.

124

125 ***Stimuli***

126 The experiment was programmed in MATLAB (MathWorks) and stimuli were presented using
127 Psychtoolbox (Brainard, 1997). Abstract, 22- or 28-line shapes were randomly created by
128 specifying an (invisible) grid of 6×6 squares that subtended 4 degrees of visual angle where lines
129 could connect two vertices horizontally, vertically, or diagonally. The first line always stemmed
130 from the central vertex of the invisible grid randomly connecting one of the surrounding eight
131 vertices to ensure shape centrality within the grid. The remaining lines were drawn sequentially,
132 ensuring all lines were connected. Orientation and originating vertices were selected randomly.

133 All words were nouns of 6 to 12 letters with 1 to 4 syllables obtained from the Medical
134 Research Council Psycholinguistic Database (Wilson, 1988). In the perceptual task, words had
135 high familiarity, concreteness, and imageability ratings (400-700). In the memory task, words
136 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 the memory task (15 mini-blocks per subject; 171 Amazon Mechanical Turk participants [73 for shapes; 98 for words] and 6 subjects in the laboratory who performed a complete version of the experiment). Based on these results, we expected a mean performance in the memory task 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).

In the perceptual task, the difference in brightness (Δb) between the two stimuli was calibrated for each subject and independently for each stimulus type. The brightness of a randomly located reference stimulus was fixed (mid-gray). The brightness of the non-reference stimulus was titrated using a staircase procedure similar to previous experiments (Fleming et al., 2010; 2012b; 2014). During practice, we used a fixed large step size 2-down/1-up procedure until subjects reached 15 reversals or 90 trials. The step sizes followed recommended ratios to

match the expected performance in memory blocks (García-Pérez, 1998). The experiment began 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.

A mid-gray fixation cross subtending 0.3 degrees of visual angle was presented between the two stimuli on a black background. 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.

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(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). Meta-d' indicates the d' that would have been predicted to give rise to the observed confidence rating data assuming a signal detection theoretic ideal observer. Meta-d'=d' indicates an optimal type II behavior for the observed type I behavior. Meta-d' greater or less than d' indicates metacognition that is better or worse, respectively, than expected given task performance, as may occur for instance if first-order decisions and confidence are supported by partly parallel processing streams (Fleming and Daw, 2017). 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)

183 from the posterior samples (Kruschke, 2010). For correlation and individual differences analyses
184 we used single-subject Bayesian model fits. Two subjects were discarded for missing more than
185 10% of the trials (i.e. >1 standard deviation from the average missed trials, which was 5%).
186 Missed trials were not analyzed.
187
188 ***fMRI data acquisition***
189 Brain images were acquired using a 3T Allegra scanner (Siemens). BOLD-sensitive functional
190 images were acquired using a T2*-weighted gradient-echo echo-planar images (42 transverse
191 slices, interleaved acquisition; TR, 2.34s; TE, 30ms; matrix size: 64×64; 3×3mm in-plane
192 resolution; slice thickness: 3mm; flip angle: 90°; FOV: 126mm). The main experiment consisted
193 of three runs of 210 volumes and three runs of 296 volumes for the perceptual and memory tasks,
194 respectively. We collected a T1-weighted MPRAGE anatomical scan (1×1×1mm voxels; 176
195 slices) and local field maps for each subject.
196
197 ***fMRI data preprocessing***
198 Imaging analysis was carried out using SPM12 (Statistical Parametric Mapping;
199 www.fil.ion.ucl.ac.uk/spm). The first five volumes of each run were discarded to allow for T1
200 stabilization. Functional images were realigned and unwarped using local field maps (Andersson
201 et al., 2001) and then slice-time corrected (Sladky et al., 2011). Each participant's structural
202 image was segmented into gray matter, white matter, cerebral spinal fluid, bone, soft tissue, and
203 air/background images using a nonlinear deformation field to map it onto template tissue
204 probability maps (Ashburner and Friston, 2005). This mapping was applied to both structural and
205 functional images to create normalized images to Montreal Neurological Institute (MNI) space.

206 Normalized images were spatially smoothed using a Gaussian kernel (8mm FWHM). We set a
207 within-run 1mm rotation and 4mm affine motion cut-off criterion, which led to the exclusion of 4
208 subjects, leaving a total of 24 subjects whose functional and behavioral data were fully analyzed.
209
210 ***Univariate analysis***
211 All our general linear models (GLMs) focus on the 'rating period' of each trial by specifying
212 boxcar regressors beginning at the subjects' type I response and ending at their type II response
213 (i.e. either confidence rating or number press). Motion correction parameters were entered as
214 covariates of no interest along with a constant term per run. Regressors were convolved with a
215 canonical hemodynamic response function (HRF). Low-frequency drifts were excluded with a
216 1/128Hz high-pass filter. Missed trials were not modeled. For judgment-related (JR) analyses,
217 we created a GLM with two regressors of interest per run to estimate BOLD response amplitudes
218 in each voxel during the 'rating period' in each trial of the Confidence and Follow blocks. For
219 the confidence level-related (CLR) parametric modulation analysis, a GLM was used to estimate
220 BOLD responses in Confidence blocks. There were two regressors of interest in each run, one
221 modeling the confidence 'rating period' and another that encoded a parametric modulation by the
222 four available confidence ratings (1-4).
223
224 *Statistical inference*
225 For the JR-analysis, single-subject contrast images of the Confidence and Follow regressors were
226 entered into a second-level random effects analysis using one-sample *t*-tests against zero to
227 assess group-level significance. For the CLR-parametric modulation analysis, single-subject
228 contrast images of the parametric modulator were entered into a similar second-level random

229 effects analysis. For conjunction analyses of activations common to both domains, second-level
230 maps thresholded at $P<0.001$ (uncorrected) were intersected to reveal regions of shared
231 statistically significant JR- and CLR-activity. Activations were visualized using MRIcro
232 (http://www.mccauslandcenter.sc.edu/crnl/mricro). All second-level unthresholded statistical
233 images were uploaded to Neurovault (Gorgolewski et al., 2015)
234 (https://neurovault.org/collections/3232/).
235
236 *ROI analysis*
237 To define regions of interest (ROIs), 12mm spheres were centered at MNI coordinates identified
238 from previous literature (Figure 3C). ROIs in left rostrolateral prefrontal cortex (L rlPFC) [-33,
239 44, 28], right rlPFC (R rlPFC) [27, 53, 25] and dorsal anterior cingulate cortex/pre-
240 supplementary motor area (dACC/pre-SMA) [0, 17, 46] were created based on (Fleming et al.,
241 2012b). The mask for precuneus (PCUN) [0, -64, 24] was based on (McCurdy et al., 2013). The
242 MNI *x*-coordinates for the dACC/pre-SMA and PCUN masks were set to 0 to ensure bilaterality.
243 Beta values were extracted from subjects' contrast images for the JR- and CLR-univariate
244 analyses, respectively.
245
246 ***Multi-voxel pattern analysis***
247 Multi-voxel pattern analysis (MVPA) was carried out in MATLAB using The Decoding Toolbox
248 (Hebart et al., 2015). We classified run-wise beta images from GLMs modeling JR- and CLR-
249 activity patterns in ROI and whole-brain searchlight analyses. ROI MVPAs were performed on
250 normalized, smoothed images using the ROI spheres as masks. Previous work has shown that
251 these preprocessing steps have minimal impact on support vector machine (SVM) classification

252 accuracy, while allowing meaningful comparison across subject-specific differences in anatomy,
253 as in standard fMRI analyses (Kamitani and Sawahata, 2010; Op de Beeck, 2010). A single
254 accuracy value per subject, per condition and per ROI was extracted and used for group analysis
255 and statistical testing. As a control, we added a 6mm-radius sphere centered at the ventricles [0 2
256 15].

257 Whole-brain searchlight analyses used 12mm-radius spheres centered around a given
258 voxel, for all voxels, on spatially realigned and slice-time corrected images from each subject to
259 create whole-brain accuracy maps. For group-level analyses, these individual searchlight maps
260 were spatially normalized and smoothed using a Gaussian kernel (8mm FWHM) and entered into
261 one-sample *t*-tests against chance accuracy (Hebart et al., 2014; 2015). Whole-brain cluster
262 inference was carried out in the same manner as in univariate analysis. Visualizations were made
263 with Surf Ice (https://www.nitrc.org/projects/surfice/).

264 Prior to decoding, for JR-activity pattern classification, we modeled two regressors of
265 interest per run focused on the 'rating periods' in the Confidence and Follow conditions. For
266 classification of CLR-activity patterns, we collapsed ratings 1 and 2 into a low-confidence
267 regressor and ratings 3 and 4 into a high-confidence regressor to allow binary classification. The
268 remaining parameters of no interest were specified as in the univariate case. For the CLR-
269 searchlight analysis, we used an exclusive mask of activity patterns associated with usage of the
270 confidence scale obtained from the successful cross-classification of button presses (1-2 vs 3-4)
271 between the Confidence and Follow conditions to eliminate low-level visuomotor confounds
272 (Figure 4D).

273 In independent across-domain classifications, we used the run-wise beta images
274 reflecting JR- and CLR-activity as pattern vectors in a linear support vector classification model

275 (as implemented in LIBSVM). We assigned each vector from each domain a label corresponding
276 to the classes Confidence (1) and Follow (-1) in the JR-analysis and Low Confidence (-1) and
277 High Confidence (1) in the CLR-analysis. We trained an SVM with the vectors from one domain
278 (3 per class, 6 in total) and tested the decoder on the 6 vectors from the other domain (and vice
279 versa) (Figure 4A; left), obtaining a mean average classification accuracy value for each of these
280 two-way cross-classifications.

281 For within-domain classifications, we ran independent leave-one-run-out cross-
282 validations for each domain on JR-activity patterns (Confidence vs Follow) and CLR-activity
283 patterns (Low vs High confidence). Pattern vectors from two of the three runs in each domain
284 were used to train an SVM to predict the same classes in the vectors from the left-out run. We
285 compared the true labels of the left-out run with the labels predicted by the model and iterated
286 this process for the other two runs to calculate a mean cross-validated accuracy independently for
287 each domain (Figure 4A; right).

288 We also tested the ability of confidence-related activity patterns to predict objective
289 performance in the absence of confidence reports. We used a GLM that modeled Low vs High
290 Confidence trials with a regressor that focused on the 'rating period', and Incorrect vs Correct
291 Follow trials with a regressor that focused on the 'decision period' (i.e. from stimulus onset to
292 subjects' type I response). We performed a cross-classification analysis in which a decoder
293 trained on Confidence trials (Low vs High Confidence) was tested on pattern vectors from
294 Follow trials (Incorrect vs Correct), and vice versa (collapsed across domain). This confidence-
295 objective performance generalization score was compared to a leave-one-run-out cross-
296 validation analysis decoding Low vs High confidence on Confidence trials only (collapsed across

297 domain). Together, these scores characterize whether a particular set of patterns are specific to
298 confidence, or also generalize to predict objective performance (Cortese et al., 2016) (Figure 5).
299
300 ***Individual differences***
301 Metacognitive efficiency scores (log meta-*d'/d'*) for each subject were estimated independently
302 for the perceptual and memory tasks, together with a single score collapsed across domains.
303 These scores were inserted as covariates in second-level analyses of within-perception, within-
304 memory, and across-domain classifications of confidence level-related activity, respectively, to
305 assess the parametric relationship between metacognitive efficiency and decoding success.
306
307 **RESULTS**
308
309 We analyzed the data of 24 subjects who underwent hemodynamic neuroimaging while
310 performing two-alternative forced-choice discrimination tasks in perceptual (PER) and memory
311 (MEM) domains (Figure 1A). In the perceptual task, subjects were asked to indicate the brighter
312 of two stimuli (words or abstract shapes). In the memory task, subjects were asked to memorize
313 exemplars of the same stimulus types, and then select the previously-learned stimulus from two
314 stimuli presented on each trial. In half of the trials ("Confidence" condition) subjects performed
315 a metacognitive evaluation after the discrimination task by rating their confidence in the
316 correctness of their response by selecting a number on a 1-to-4 scale (1=not confident; 4=very
317 confident). In order to differentiate metacognitive judgment-related activity from visuomotor
318 activity engaged by use of the confidence scale, in the other half of trials ("Follow" condition)
319 subjects were asked to respond according to a highlighted number without evaluating confidence

in their response. To avoid stimulus-type confounds, two different types of stimulus—words and abstract shapes—were used as stimuli in both tasks.

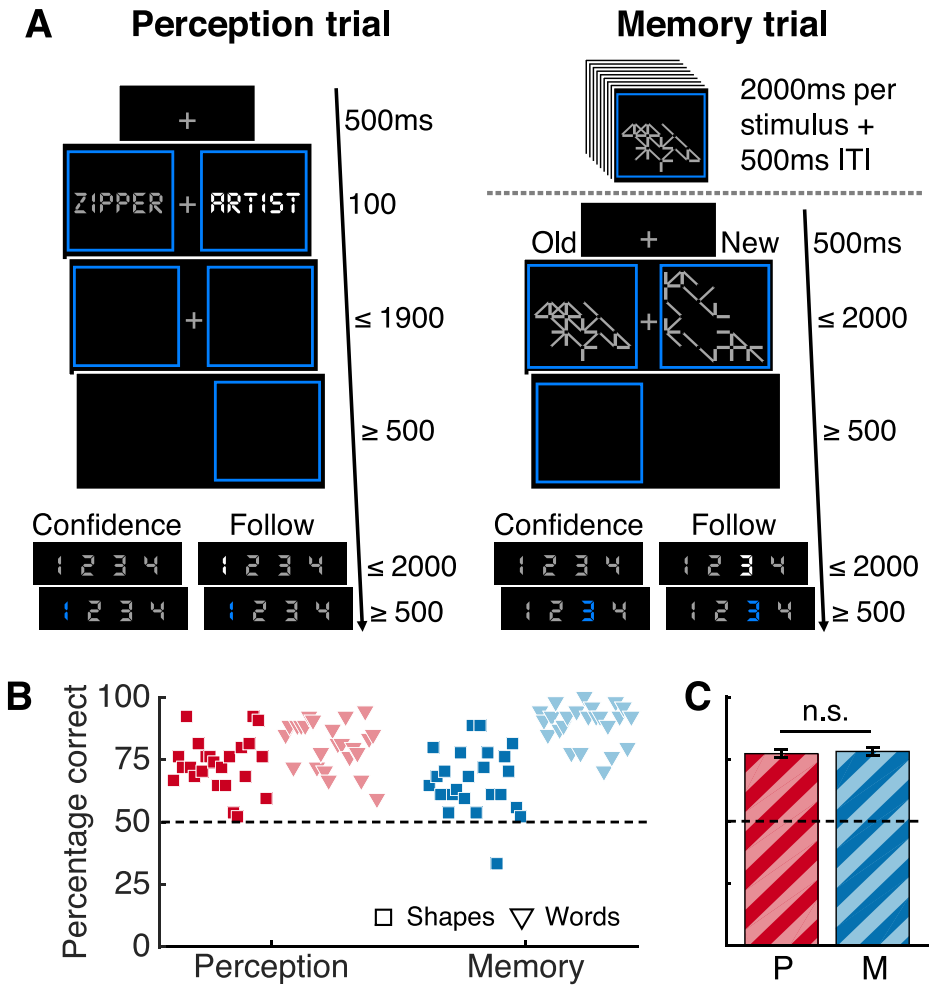


Figure 1. Task design & performance results. (A) Subjects performed two-alternative forced-choice discrimination tasks about perception and memory. In perception blocks, subjects selected the brighter of two stimuli. Memory blocks started with an encoding period and then subjects indicated in each trial 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)** 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

336 ***Behavior***

337 We first compared task performance, measured by percentage of correct responses, across
338 condition, task, and stimulus type. A 2×2×2 repeated measures ANOVA (CONFIDENCE/FOLLOW ×
339 PERCEPTION/MEMORY × SHAPES/WORDS) showed that performance was well-matched across
340 conditions (Confidence vs Follow) ($F_{1,23}$=3.036, P=0.095). None of the four paired t-tests
341 (DOMAIN×STIMULUS) comparing performance between the Confidence and Follow conditions
342 returned a significant difference (P>0.05). In the remainder of the behavioral analyses we focus
343 on the Confidence condition. Matching performance across stimulus type was more challenging
344 because subjects' memory for words was expected to be considerably higher than that for
345 abstract shapes trials based on pilot data (see Methods for details). Instead, we aimed to match
346 subjects' performance independently for each stimulus type across task domains by titrating the
347 difficulty of the perceptual task to approximate the performance expected for the corresponding
348 stimulus type in the memory task (shapes: PER M=73%, MEM M=67%; words: PER M=81%, MEM
349 M=89%; Figure 1B). Critically, this ensured performance was matched across task domains
350 when averaging stimulus types across participants (PER: M=77%, MEM: M=78%; paired t-test
351 T_{23}=0.38, P=0.70; Figure 1C). A 2×2 repeated measures ANOVA of performance in the
352 Confidence condition (PERCEPTION/MEMORY × SHAPES/WORDS) confirmed there was no main
353 effect of domain ($F_{1,23}$= 0.15, P=0.702). However, we observed a main effect of stimulus type
354 due to greater overall performance on words ($F_{1,23}$=75.69, P= 9.87×10^{-9}) and a DOMAIN ×
355 STIMULUS interaction, due to a greater difference in performance between shapes and words in
356 the memory compared to the perception task ($F_{1,23}$=16.74, P=0.00045).

357 Subjects were faster providing type I responses in perceptual trials (M=636ms) than in
358 memory trials (M=1222ms). There was also a small difference in reaction times between shape

(M=967ms) and word (M=892ms) trials. A 2×2 repeated measures ANOVA confirmed a main effect of domain ($F_{1,23}$=367, P=1.23×10^{-15}) driven by slower reaction types in the memory task. There was also a main effect of stimulus type on response time ($F_{1,23}$=8.95, P=0.006), as well as a significant DOMAIN × STIMULUS interaction ($F_{1,23}$=5.82, P=0.024).

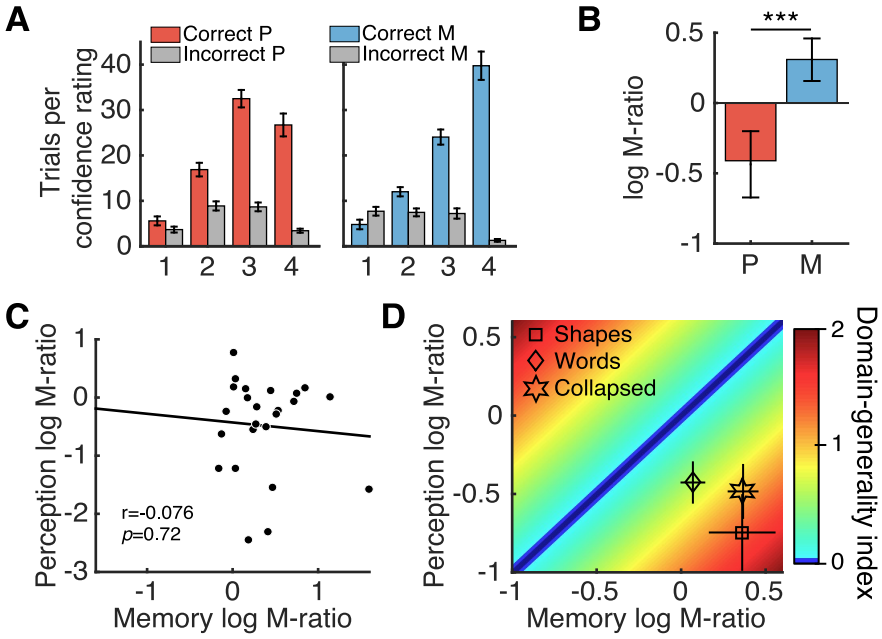


Figure 2. Metacognitive measures. (A) Mean number of correct and incorrect trials per confidence rating. **(B)** Metacognitive efficiency measured by log(meta-*d*'/*d*'). Zero indicates that metacognitive sensitivity (meta-*d*') is equal to task sensitivity *d*' (i.e. the *d*' that would have been predicted 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)** Domain-general index (DGI) for each subject that quantifies the similarity between their metacognitive efficiency scores in each domain (see main text for details). Greater DGI scores indicate less metacognitive consistency across domains. Mean log(meta-*d'/d'*) values for each stimulus type in both domains 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

380 (PER M=2.62, MEM M=2.47; paired *t*-test T_{23}=1.26, *P*=0.22). Reaction times for confidence
381 ratings were not different between domains (PER M=518ms, MEM M=516ms; paired *t*-test
382 T_{23}=0.16, *P*=0.87). We next estimated log (meta-*d'*/*d'*), a metacognitive efficiency measure
383 derived from signal detection theory that assays the degree to which confidence ratings
384 distinguish between correct and incorrect trials (Maniscalco and Lau, 2012; Fleming and Lau,
385 2014; Fleming, 2017). We used hierarchical Bayesian estimation to incorporate subject-level
386 uncertainty in group-level parameter estimates (Fleming, 2017). Metacognitive efficiency in the
387 perceptual task was significantly lower than in the memory task ($P_{\theta>0}$ ~ 1; see Figure 2B &
388 Methods for details), consistent with previous findings (Fleming et al., 2014). Metacognitive
389 efficiency above optimality (meta-*d'*=*d'*) in memory trials suggests subjects had better
390 metacognition than expected given their task performance, while suboptimal metacognitive
391 efficiency in perceptual trials suggests subjects had worse metacognition than expected given
392 their task performance (assuming an ideal observer in both cases). We did not find a correlation
393 between subjects' individual metacognitive efficiency scores in the perceptual and memory
394 domains (r_{22}=-0.076; *P*=0.72; Figure 2C). We also evaluated the correlation coefficient within a
395 hierarchical model of meta-*d'*, which takes into account uncertainty in subject-level model fits
396 (Fleming, 2017). The 95% confidence interval on the posterior correlation coefficient overlapped
397 zero in this analysis (ρ=0.205; HDI=[0.826, -0.358]), also indicating a dissociation between
398 domains.

399 We next estimated metacognitive efficiency separately for each stimulus type (Figure
400 2D). A 2×2 repeated measures ANOVA (PERCEPTION/MEMORY × SHAPES/WORDS) indicated
401 metacognitive efficiency was greater for memory than perception ($F_{1,23}$=22.44, *P*=8.97×$^{10-5}$).
402 Importantly, there was no stimulus main effect ($F_{1,23}$=0.015, *P*=0.902) and there was no

403 interaction between domain and stimulus type ($F_{1,23}$=2.835, P=0.106). To further assess a
404 potential covariation between metacognitive abilities in each domain, we calculated for each
405 subject a domain-generality index (DGI) that quantifies the similarity between scores in each
406 domain for each participant (Fleming et al., 2014):
407
408 $$\text{DGI} = |\log M_P - \log M_M|$$
409
410 where M_P = perceptual meta-*d'/d'* and M_M = memory meta-*d'/d'*. Lower DGI scores indicate
411 more similar metacognitive efficiencies between domains (DGI=0 indicates identical scores).
412 Mean DGI for shapes (1.42), words (0.66) and collapsed by stimulus type (0.95) were higher
413 than zero (Figure 2D). Metacognition for words was behaviorally more stable across domains as
414 the DGI was smaller than for shapes (paired *t*-test: T_{23}=2.86; P=0.009). Together, these results
415 suggest domain-specific constraints on metacognitive ability.
416
417 ***fMRI analyses***
418
419 We next turned to our fMRI data to assess the overlap between neural substrates engaged when
420 metacognitive judgments are made during perceptual and memory tasks. A full understanding of
421 the neural substrates of metacognition requires an independent examination of the process of
422 engaging in a metacognitive task and the level of confidence expressed by the subject (Chua et
423 al., 2014). To this end, in both univariate and multivariate analyses, we focused on two distinct
424 features of metacognition-related activity. First, we assessed brain regions engaged in judgment-
425 related (JR) activity (i.e. the difference between Confidence trials requiring a metacognitive

426 judgment and the Follow condition). Second, we assessed brain regions engaged in confidence
427 level-related (CLR) activity. In univariate CLR-analysis, we focused on the parametric
428 relationship between confidence ratings (1 through 4) and neural activity. In multivariate CLR-
429 analysis, we collapsed ratings 1 and 2 into a low-confidence category and ratings 3 and 4 into a
430 high-confidence category to allow binary classification of activity patterns.
431
432 *Univariate results*
433 *Judgment-related activity.* In standard univariate analyses, we found elevated activity in
434 dACC/pre-SMA, bilateral insulae and superior and middle frontal gyri when contrasting the
435 Confidence against the Follow condition (collapsed by domain), consistent with previous
436 findings (Fleming et al., 2012b) (Figure 3A). There were no significant clusters of activity in the
437 reverse contrast (Follow>Confidence). Splitting the data by domain (see Table 1), an interaction
438 contrast [Memory Confidence > Memory Follow] > [Perception Confidence > Perception
439 Follow] revealed significant clusters of activity in middle cingulate gyrus, left insula, precuneus,
440 left hippocampus and cerebellum (Figure 3B, blue). No significant clusters of activity were
441 found in the reverse interaction contrast. In a conjunction analysis, elevated activity for the
442 Confidence > Follow condition was observed across both perception and memory trials in
443 anterior cingulate and right insula (Figure 3B, green).

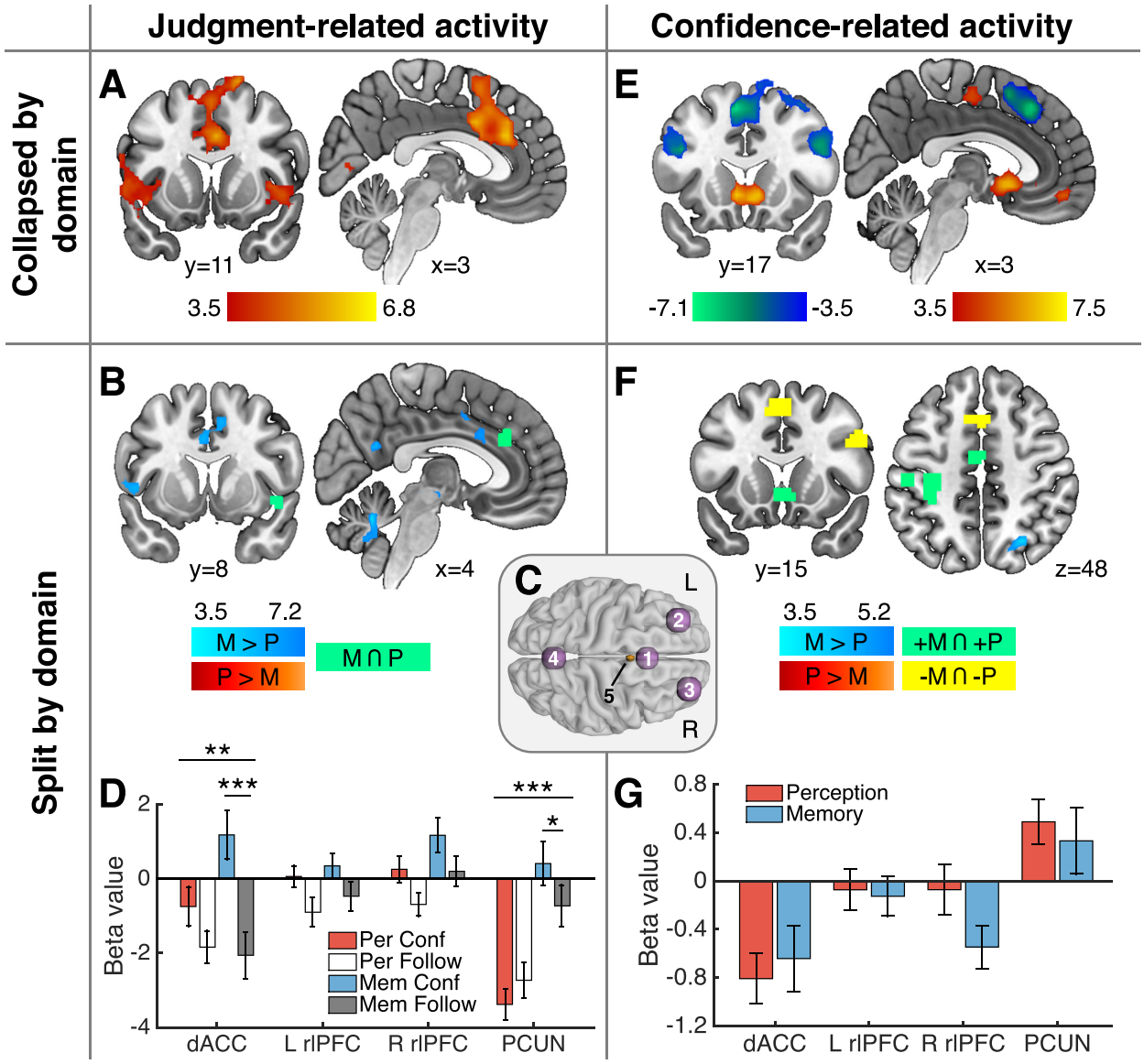


Figure 3. fMRI univariate analysis results. *Judgment-related activity*: **(A)** Whole-brain analysis of significant activation in the Confidence>Follow contrast (collapsed by domain); there were no significant clusters in the Follow>Confidence contrast. **(B)** [Memory Confidence > Memory Follow] > [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. **(C)** Spherical binary masks of four a priori regions of interest [1=dACC, 2=L rlPFC, 3=R rlPFC, 4=PCUN] and an ROI in the ventricles [5] used as a control region in multivariate analyses (Figure 4). **(D)** Estimated mean beta values for judgment-related activity by domain in the main four ROIs displayed in (C). *Confidence level-related activity*: **(E)** Whole-brain analysis of activity parametrically modulated by level of confidence (collapsed by domain). 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

459 conjunction analysis revealed shared activity that was positively (green) and negatively (yellow)
460 correlated with confidence levels in both domains. **(G)** Estimated mean beta values of confidence
461 level-related activity in the main four ROIs displayed in (C). All displayed whole-brain
462 activations are significant at a cluster-defining threshold $P<0.001$, corrected for multiple
463 comparisons $P_{FWE}<0.05$; except for conjunction analyses in which we computed the intersection
464 of two independent maps thresholded at $P<0.001$, uncorrected. Images displayed at $P<0.001$.
465 Graded color bars reflect T-statistics. Error bars indicate s.e.m. *** $P<0.001$ ** $P<0.01$ *$P<0.05$.
466 L=left; R=right; P=perception; M=memory.
467

468 To further quantify these effects for each task domain, we focused on four a priori
469 regions of interest (ROI) in (dACC/pre-SMA), bilateral rostrolateral prefrontal cortex (rlPFC)
470 and precuneus (PCUN) (see Figure 3C & Methods), which previous studies have found to be
471 recruited by perceptual and memory metacognition (Fleck et al., 2006; Fleming et al., 2010;
472 2012b; Baird et al., 2013; McCurdy et al., 2013). In a series of repeated measures 2×2 ANOVAS
473 (CONDITION × TASK) we found a main effect of greater activity on Confidence compared to
474 Follow trials in all ROIs except precuneus, where instead we observed a main effect of task, with
475 increased activity on memory trials (Figure 3D; condition main effect: dACC/pre-SMA,
476 $F_{1,23}=19.34$, $P=0.0002$; left rlPFC, $F_{1,23}=6.62$, $P=0.017$; right rlPFC, $F_{1,23}=9.28$, $P=0.006$; PCUN:
477 $F_{1,23}=0.40$, $P=0.532$; task main effect: dACC/pre-SMA, $F_{1,23}=2.33$, $P=0.14$; left rlPFC,
478 $F_{1,23}=0.95$, $P=0.34$; right rlPFC, $F_{1,23}=4.94$, $P=0.036$; PCUN: $F_{1,23}=36.78$, $P=3.47\times10^{-6}$). We
479 found that the difference between Confidence and Follow trials was greater in memory than in
480 perception trials in dACC/pre-SMA, recapitulating the whole-brain results (CONDITION × TASK
481 interaction; $F_{1,23}=12.16$, $P=0.0019$; paired *t*-test, MEM: $T_{23}=5.47$, $P=0.0001$, PER: $T_{23}=1.92$, $P=$
482 0.067). A similar interaction pattern was observed in precuneus (CONDITION × TASK interaction:
483 $F_{1,23}=15.86$, $P=0.0006$; paired *t*-test, MEM: $T_{23}=2.43$, $P=0.023$, PER: $T_{23}=-1.54$, $P=0.136$). There
484 were no interactions in frontal regions (left rlPFC, $F_{1,23}=0.07$, $P=0.795$; right rlPFC, $F_{1,23}=0.002$,

485 *P*=0.968). These results are compatible with previous findings indicating a distinctive
486 contribution of precuneus to memory metacognition (Baird et al., 2013; McCurdy et al., 2013).
487
488 *Confidence level-related activity.* We next sought to investigate the parametric relationship
489 between confidence level and neural activity. Collapsing across domains, we found activity in
490 the left pre- and post-central gyri, the posterior midline, ventral striatum and ventromedial
491 prefrontal cortex (vmPFC) correlated positively with confidence ratings (Figure 3E, hot colors).
492 We also replicated negative correlations between confidence and activation in dACC/pre-SMA,
493 parietal cortex, and bilateral PFC that have been reported in several previous studies (Fleck et al.,
494 2006; Fleming et al., 2012b; Baird et al., 2013; Hebart et al., 2014) (Figure 3E, cool colors).
495 When testing for differences between these parametric regressors by domain (see Table 2), a
496 Memory > Perception contrast revealed a significant cluster of activity in right parietal cortex
497 (Figure 3F, blue), while there was no significant activity in a Perception > Memory contrast.
498 Shared positive correlations between confidence and activity in perception and memory trials
499 were found in ventral striatum and in left pre- and post-central gyri, the latter consistent with use
500 of the right hand to provide confidence ratings (conjunction analysis; Figure 3F, green). Shared
501 negative correlations with confidence were found in regions of right dorsolateral PFC and medial
502 prefrontal cortex, overlapping with pre-SMA (Figure 3F, yellow).
503 Complementing the ROI analysis of JR-activity, we performed an ROI analysis of CLR-
504 activity that recapitulated the whole-brain results. We observed negative relationships between
505 confidence and activity in dACC/pre-SMA and positive relationships in precuneus. Importantly,
506 no significant differences in the parametric effect of confidence were found between domains in
507 any of our a priori ROIs (Figure 3G; paired *t*-tests: dACC/pre-SMA, T_{23}=-0.47, *P*=0.643; left

508 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).
509 Together with the lack of marked domain-specific differences in confidence-related activity at
510 the whole-brain level, these results are suggestive of an absence of domain-specificity in
511 confidence-related activity. However, a lack of difference between univariate activation profiles
512 is not necessarily conclusive. For instance, differences in confidence level may be encoded in
513 fine-grained spatial patterns of activity even when the overall BOLD-activity is evenly
514 distributed across confidence levels (Cortese et al., 2016). Similarly, while metacognition-related
515 activity may show similar overall levels of activation across tasks, distributed activity patterns in
516 frontal and parietal areas may carry distinct task-specific information (Hebart et al., 2014; Cole
517 et al., 2016). We next turned to multivariate analysis methods, which are sensitive to differences
518 in spatial activity patterns, to test this hypothesis.
519
520 *Multivariate results*
521 We performed a series of multi-voxel pattern analyses (MVPA) (Figure 4A) focused on both
522 judgment-related activity patterns and confidence level-related activity patterns.
523
524 *ROI analysis of judgment-related activity patterns.* If metacognitive judgments are based on
525 domain-general processes (i.e. shared across perceptual and memory tasks), a decoder trained to
526 classify JR-activity patterns in perceptual trials should accurately discriminate JR-activity
527 patterns when tested on memory trials (and vice versa). Alternatively, domain-specific activity
528 profiles would be indicated by significant within-domain classification of JR-activity patterns in
529 the absence of across-domain transfer. To adjudicate between these hypotheses, we performed a
530 support vector machine (SVM) decoding analysis using as input vectors the run-wise beta

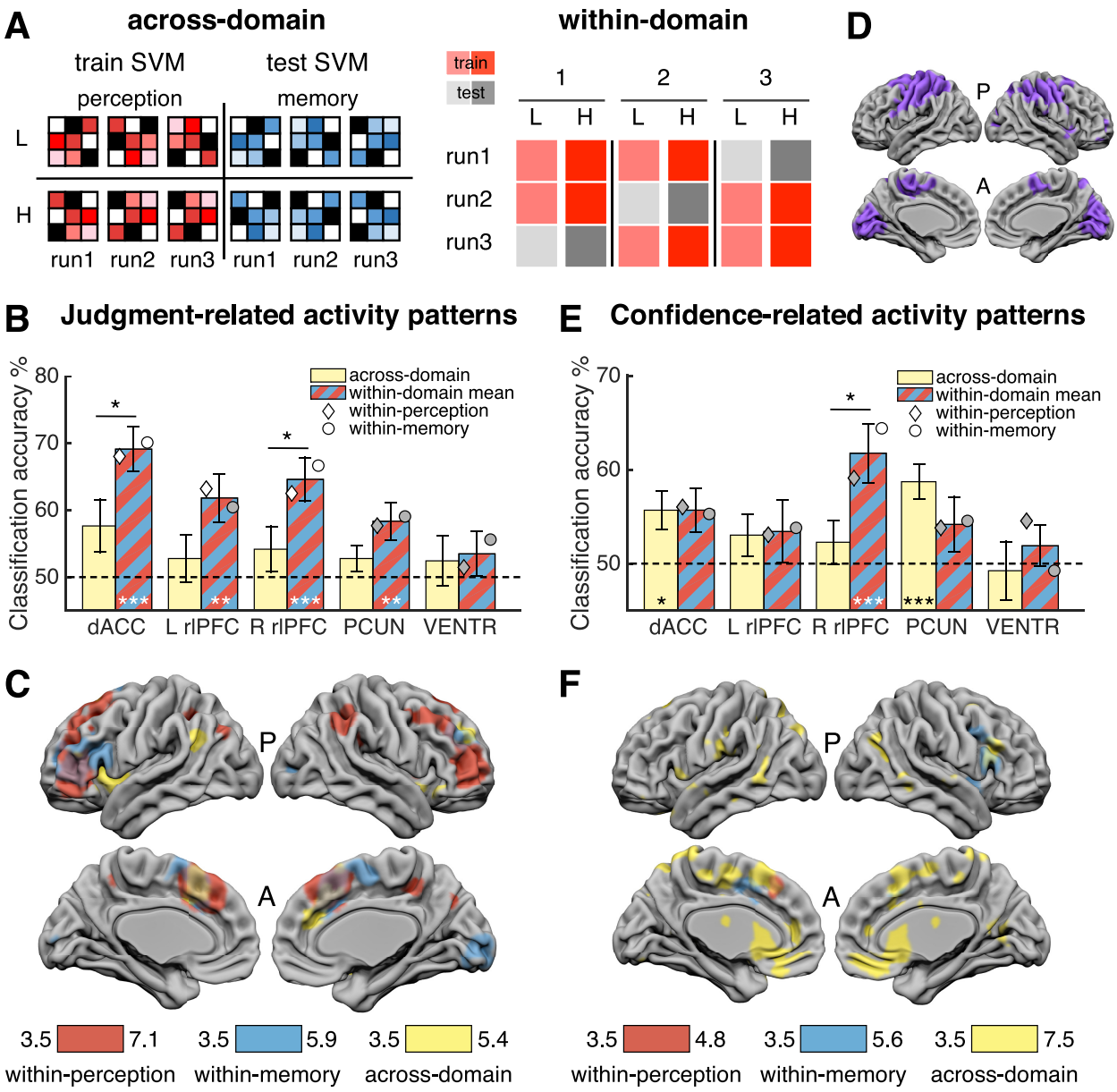
531 images pertaining to Confidence and Follow trials obtained from a GLM (12 input vectors in

532 total). For within-domain classification we used standard leave-one-out independent cross-

533 validations for each domain and we tested for across-domain generalization using a cross-

534 classification analysis (see Methods for details). Chance classification in both analyses was 50%.

535



536

537 **Figure 4. MVPA Results.** *Classification designs*: **(A)** *Left*: Across-domain classification design.
538 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). Classification of low (L) and high (H) confidence levels is illustrated here. *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 process 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 Patterns*: **(B)** ROI results for across-domain (yellow) and mean within-domain (red-blue stripe) classification accuracy of Confidence vs Follow trials. **(C)** Searchlight analysis for same classifications as in (B). **(D)** Low-level visuomotor mask used in (F) (see main text and Methods for details). *Confidence Level-Related Activity Patterns*: **(E)** Low versus high confidence classification accuracy results. **(F)** Searchlight analysis for the same classifications as in (E) exclusively masked for visuomotor-related activity patterns. Bars in (B) and (E) indicate means and error bars s.e.m. Dashed lines indicate chance classification (50%). Diamonds and circles indicate mean independent classification in perception and memory trials, respectively. White diamonds/circles indicate classification was significantly different from chance, Bonferroni corrected. All clusters in (C) and (F) are significant at cluster-defining threshold $P<0.001$, corrected for multiple comparisons at $P_{FWE}<0.05$. Image displayed at $P<0.001$. Color bars indicate T-scores. P=perception; M=memory; A=anterior; P=posterior. *** $P\leq0.001$ ** $P\leq0.01$ * $P<0.05$; all one-sample t-tests are Bonferroni corrected.

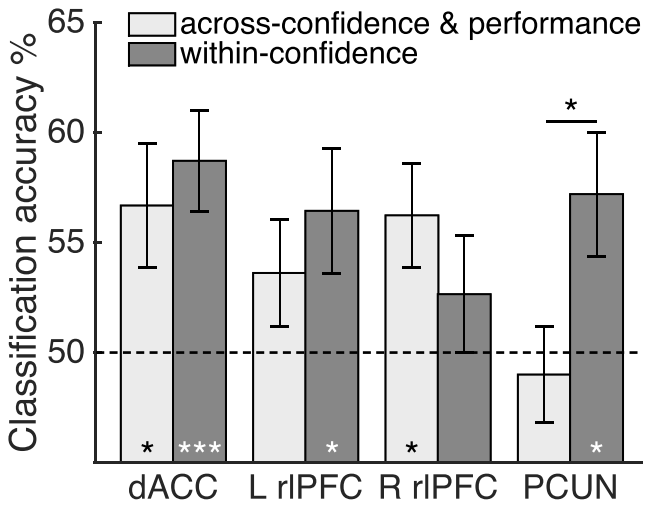
Mean within-domain classification results were significantly above chance in all ROIs (one-sample *t*-tests Bonferroni corrected for multiple comparisons $\alpha=0.05/4=0.0125$: dACC/pre-SMA $T_{23}=5.77$, $P=6.99\times10^{-6}$; L rlPFC $T_{23}=3.27$, $P=0.003$; R rlPFC $T_{23}=4.47$, $P=0.0002$; PCUN $T_{23}=2.98$, $P=0.007$; Figure 4B, red-blue stripe bars). In contrast, across-domain generalizations were not significantly different from chance in any ROI (one sample *t*-test Bonferroni corrected: dACC/pre-SMA, $T_{23}=1.95$, $P=0.06$; L rlPFC, $T_{23}=0.79$, $P=0.44$; R rlPFC, $T_{23}=1.24$, $P=0.23$; PCUN $T_{23}=1.40$, $P=0.17$; Figure 4B, yellow bars). As a control, classification accuracy in the ventricles (VENTR) was not different from chance (across: $T_{23}=0.66$, $P=0.52$; within: $T_{23}=1.04$, $P=0.31$). This suggests that the patterns of activity that distinguish metacognitive judgments from the visuomotor control condition in one domain are distinct from analogous patterns in the other domain. In particular, within-domain classification accuracy was significantly different from across-domain classification accuracy in (dACC/pre-SMA: $T_{23}=2.88$, $P=0.008$) and right

572 rlPFC (T_{23}=2.24, P=0.035). These results are consistent with the hypothesis that metacognitive
573 judgments recruit domain-specific patterns of cortical activity in prefrontal cortex.
574
575 *Searchlight analysis of judgment-related activity patterns.* We ran a similar decoding analysis
576 using an exploratory whole-brain searchlight, obtaining a classification accuracy value per voxel
577 (Hebart et al., 2015). Consistent with our ROI results, we observed significant within-domain
578 classification in large swathes of bilateral PFC for both perception (red) and memory (blue)
579 (Figure 4C). Within-perception classification was also successful in parietal regions—precuneus
580 in particular—and within-memory activity patterns were classified accurately in occipital
581 regions. We also identified clusters showing significant across-domain generalization (yellow) in
582 dACC, pre-SMA, SFG (BA 9), supramarginal gyrus (BA 40), and bilateral IFG/insula, consistent
583 with univariate results (Figure 3A).
584
585 *ROI analysis of confidence level-related activity patterns.* We next asked whether confidence is
586 encoded in a domain-general or domain-specific fashion by applying a similar approach to
587 discriminate low versus high confidence trials. Note that in this case, ROI univariate analyses did
588 not reveal any differences in confidence-related activity between domains (Figure 3G). We
589 hypothesized that if confidence level is encoded by domain-general neural activity patterns, it
590 should be possible to train a decoder to discriminate low (1-2) from high (3-4) confidence rating
591 patterns in the perceptual task and then accurately classify confidence patterns on the memory
592 task (and vice versa). In the absence of across-domain classification, significant within-domain
593 classification is indicative of confidence level-related domain-specific activity patterns. ROI
594 cross-classifications and cross-validations were performed in a similar fashion as above. We note

595 that two subjects did not provide ratings for one of the classes in at least one run and were left
596 out from the main analysis to avoid entering unbalanced training data into the classifier;
597 however, including those subjects did not change the main result.
598 One-sample *t*-tests (Bonferroni-corrected) showed across-domain classification of
599 confidence was significantly above chance in dACC/pre-SMA (T_{21}=2.83, P=0.010) and
600 precuneus (T_{21}=4.69, P=0.0001), indicative of a generic confidence signal, but not in rlPFC (left:
601 T_{21}=1.36, P=0.19; right: T_{21}=0.97, P=0.34) (Figure 4E, yellow bars). In contrast, mean within-
602 domain classification accuracy was significantly above chance in right rlPFC (T_{21}=3.75,
603 P=0.001), but not in the other ROIs (dACC/pre-SMA: T_{21}=2.42, P=0.025; left rlPFC: T_{21}=1.03,
604 P=0.32; PCUN: T_{21}=1.42, P=0.17; Bonferroni-corrected; Figure 4E, red-blue stripe bars). The
605 mean confidence classification accuracy in right rlPFC was 62% (PER=59%, MEM=64%),
606 notably above a recently estimated median 55% for decoding task-relevant information in frontal
607 regions (Bhandari et al., 2017). Importantly, classification accuracy in this ROI also differed
608 from the corresponding across-domain classification accuracy (paired *t*-test T_{21}=2.37, P=0.028).
609 Classification accuracy in the ventricles was not different from chance (across: T_{21}=-0.24,
610 P=0.81; within: T_{21}=0.86, P=0.40). When subjects with unbalanced data were included, within-
611 domain classification accuracy in right rlPFC remained at 62%, significantly above chance (T_{23}=
612 4.22; P=0.0003) and significantly different from across-domain classification accuracy (paired t-
613 test: T_{23}= 2.54; P=0.018). Together, these results suggest the co-existence of two kinds of CLR-
614 neural activity: dACC/pre-SMA and precuneus encode a generic confidence signal, whereas
615 patterns of activity in right rlPFC were modulated by task, allowing within- but not across-
616 domain classification of confidence level.
617

618 *Searchlight analysis of confidence level-related activity patterns.* We ran a similar decoding
619 analysis of confidence level using an exploratory whole-brain searchlight, obtaining a
620 classification accuracy value per voxel. Here we leveraged the Follow trials as a control for low-
621 level visuomotor confounds by exclusively masking out activity patterns associated with usage
622 of the confidence scale (Figure 4D). The remaining activity patterns can therefore be ascribed to
623 confidence level-related signals that do not encode visual or motor features of the rating (Figure
624 4F). We found widespread across-domain classification of confidence (yellow) in a
625 predominantly midline network including a large cluster encompassing dACC/pre-SMA,
626 vmPFC, and ventral striatum. Domain-specific confidence level-related activity patterns were
627 successfully decoded from right PFC (insula, IFG, BA 9, 46) in memory trials (blue) and were
628 also independently decoded in both domains from dACC/pre-SMA.
629
630 *Generalization of CLR-activity to objective performance.* To further address the question of how
631 confidence judgments may relate to activity patterns, we examined the relationship between
632 objective task accuracy and confidence. Previous work suggests that the neural basis (and
633 associated activation patterns) of confidence and performance may be partly distinct (Rounis et
634 al., 2010; Cortese et al., 2016). Specifically, we tested the hypothesis that we could train a
635 decoder using CLR-activity patterns to classify objective performance-related activity patterns
636 (correct/incorrect) on Follow trials (and vice versa) in a cross-classification analysis (collapsed
637 across domain). This analysis confirmed that activity patterns in dACC/pre-SMA (T_{21}=2.38,
638 P=0.027) and right rlPFC (T_{21}=2.64, P=0.015) could predict objective accuracy levels in Follow
639 trials above chance (Figure 5, light gray; uncorrected), but not in left rlPFC (T_{21}=1.49, P=0.15)
640 or precuneus (T_{21}=-0.46, P=0.65). We then compared these decoding scores to a leave-one-run-

641 out cross-validation decoding analysis of low versus high confidence on Confidence trials only
642 (collapsed by domain; Figure 5, dark gray; uncorrected). Consistent with the analyses reported in
643 Figure 4E (yellow), this decoder was unable to classify domain-general confidence patterns of
644 activity in right rlPFC (T_{21}=1.00; P=0.33), but was above chance in dACC/pre-SMA (T_{21}=3.80,
645 P=0.001), left rlPFC (T_{21}=2.26, P=0.034) and precuneus (T_{21}=2.56, P=0.018). Critically, in
646 precuneus, confidence classification was significantly greater than confidence-performance
647 generalization, which was at chance (paired t-test T_{21}=2.16, P=0.043). This result indicates that
648 confidence-related patterns in precuneus do not generalize to predict objective performance,
649 consistent with a partly distinct coding of information relevant to task performance and
650 confidence. In contrast, in dACC/pre-SMA, general confidence level and performance could be
651 predicted from common patterns of activation.
652



653

654 **Figure 5. Generalization of confidence-related activity to objective accuracy.** Light gray bars
655 denote mean cross-classification accuracy results obtained from training a decoder on CLR-
656 activity and testing it on objective accuracy (correct / incorrect) activity patterns in the Follow
657 condition (and vice versa). Dark gray bars denote decoding accuracy for a leave-one-out cross-
658 validation of low and high confidence on Confidence trials only (collapsed by domain). Bars
659 indicate group means and bars s.e.m. Dotted line indicates chance level. * P<0.05 *** P<0.001
660

661 *Metacognitive efficiency and confidence level-related activity classification.* Finally, we
662 reasoned that if confidence-related patterns of activation are contributing to metacognitive
663 judgments, they may also track individual differences in metacognitive efficiency. To test for
664 such a relation, we asked whether individual metacognitive efficiency scores collapsed across-
665 domains and independently in each domain predicted searchlight classification accuracy of
666 confidence level. We did not find any significant clusters after whole-brain correction for
667 multiple comparisons in domain-general, within-perception, or within-memory analyses.
668 However, memory metacognitive efficiency predicted memory confidence classification
669 accuracy in a cluster in right precuneus and left precentral gyrus ($P<0.001$, uncorrected), whereas
670 perceptual metacognitive efficiency predicted perceptual confidence classification accuracy in
671 left middle frontal gyrus, right vmPFC, bilateral temporal gyri, and cerebellum ($P<0.001$,
672 uncorrected). Previous studies (Fleming et al., 2010; McCurdy et al., 2013; Fleming et al., 2014)
673 have reported similar relations between perceptual and memory metacognitive efficiency and
674 individual differences in the structure of prefrontal and parietal cortex, respectively. While we do
675 not interpret these findings further in the current manuscript, for completeness, second-level
676 unthresholded statistical images of these analyses were uploaded to Neurovault in order to
677 inform future work (Gorgolewski et al., 2015) (https://neurovault.org/collections/3232/).

678 **DISCUSSION**

679 When performing a cognitive task, confidence estimates allow for comparisons of performance
680 across a range of different scenarios (de Gardelle and Mamassian, 2014). Such estimates must
681 also carry information about the task context if they are to be used in decision-making. Here we
682 investigated the domain-generality and domain-specificity of representations that support
683 metacognition of perception and memory.

684 Unlike previous studies (McCurdy et al., 2013), subjects' performance was matched
685 between domains for two different types of stimulus, thereby eliminating potential performance
686 and stimulus confounds. Subjects' confidence ratings were also matched between domains and
687 followed expected patterns of higher ratings after correct decisions than after incorrect decisions.
688 Metacognitive efficiency scores between tasks were not correlated, and metacognitive efficiency
689 scores in the memory task were superior to those in the perceptual task. Using univariate and
690 multivariate analyses, we showed the existence of both domain-specific and domain-general
691 metacognition-related activity during perceptual and memory tasks. We report four main
692 findings, and discuss each of these in turn.

693 First, we obtained convergent evidence from both univariate and multivariate analyses
694 that a cingulo-opercular network centered on dACC/pre-SMA encodes a generic signal
695 predictive of confidence level and objective accuracy across memory and perceptual tasks.
696 Previous studies of metacognition have implicated the cingulo-opercular network in tracking
697 confidence level (Fleck et al., 2006; Fleming et al., 2012b; Hebart et al., 2014; Hilgenstock et al.,
698 2014). However, we go beyond these previous studies to provide evidence that these signals
699 generalize to predict confidence across two distinct cognitive domains. This finding is consistent
700 with posterior medial frontal cortex as a nexus for monitoring the fidelity of generic

701 sensorimotor mappings, building on previous findings that error-related event-related potentials
702 originating from this region are sensitive to variation in subjective certainty (Scheffers and
703 Coles, 2000; Boldt and Yeung, 2015). The activity in dACC/pre-SMA was also consistently
704 elevated by the requirement for a metacognitive judgment (Fleming et al., 2012b). However, the
705 results regarding the generalizability of the pattern of these increases across tasks were
706 inconclusive. Whole-brain searchlight analysis revealed successful cross-classification of these
707 activity patterns in dACC and insular regions, consistent with the results of the univariate
708 analysis. These patterns of activity, however, did not generalize across tasks in a pre-defined
709 region of interest centered in dACC/pre-SMA.

710 Notably, while both dACC/pre-SMA and precuneus showed significant domain-general
711 decoding of confidence, in precuneus these patterns did not generalize to also predict changes in
712 objective accuracy. While performance and subjective confidence may both depend on similar
713 decision variables (Kiani and Shadlen, 2009; Fleming and Daw, 2017), behavioral dissociations
714 between these quantities are also consistent with distinct internal states contributing to decisions
715 and confidence ratings (Busey et al., 2010; Fleming and Daw, 2017). For instance, hierarchical
716 models of confidence formation suggest a downstream network "reads out" decision-related
717 information in a distinct neural population (Insabato et al., 2010). The observed lack of cross-
718 classification in precuneus (Figure 5) is consistent with the recent observation of distinct neural
719 patterns of activity pertaining to confidence and first-order performance revealed through multi-
720 voxel neurofeedback in frontal and parietal regions (Cortese et al., 2016).

721 Second, in lateral anterior frontal cortex we found activity patterns that tracked both the
722 requirement for metacognitive judgments and level of confidence. Large swathes of lateral
723 prefrontal cortex distinguished activity patterns pertaining to metacognitive judgments that were

724 specific for each domain. Critically, however, confidence-related activity patterns were selective
725 for domain in right rlPFC (Figure 4E): they differed according to whether the subject was
726 engaged in rating confidence about perception or memory. Such signals may support the
727 "tagging" of confidence with contextual information, thereby facilitating the use of confidence
728 for behavioral control (Donoso et al., 2014; Purcell and Kiani, 2016). The identity of perceptual
729 and memory tasks can be reliably decoded from activity in right PFC neural populations
730 (Mendoza-Halliday and Martinez-Trujillo, 2017), consistent with the possibility that this
731 contextual information is recruited during confidence rating. It is possible that anterior prefrontal
732 regions combine generic confidence signals with domain-specific information to fine-tune
733 decision-making and action selection in situations in which subjects need to regularly switch
734 between tasks or strategies on the basis of their reliability (Donoso et al., 2014). An alternative
735 hypothesis, also compatible with our data, is that PFC first estimates the confidence level
736 specifically for the current task, which is then relayed to medial areas to recruit the appropriate
737 resources for cognitive control in a task-independent manner. Processing dynamics may also
738 unfold simultaneously in both areas. These possibilities echo a longstanding debate in the
739 cognitive control literature on the relative primacy of medial and lateral PFC in the hierarchy of
740 control (Kerns et al., 2004; Tang et al., 2016). Further inquiry and development of computational
741 models of the hierarchical or parallel functional coupling of these networks in metacognition is
742 necessary.

743 Third, we obtained convergent evidence that precuneus plays a specific role in
744 metamemory judgments. In univariate fMRI analyses, we found the requirement for a
745 metacognitive judgment recruited our pre-established region of interest centered on precuneus
746 only on memory, but not perceptual, trials (Figure 3D). Individual metacognitive efficiency

747 scores in memory trials predicted classification accuracy in a more dorsal precuneal region,
748 while individual differences in metacognitive efficiency scores in perceptual trials predicted
749 classification accuracy in ventromedial prefrontal cortex (albeit at uncorrected thresholds). These
750 findings are consistent with the medial parietal cortex making a disproportional contribution to
751 memory metacognition (Simons et al., 2010; Baird et al., 2013; McCurdy et al., 2013), and offer
752 a potential explanation for a decrease in perceptual, but not memory, metacognitive efficiency
753 seen in patients with frontal lesions (Fleming et al., 2014). However, we do not wish to conclude
754 that precuneus involvement is specific to metamemory. We note that univariate negative
755 correlations with confidence were found also on perceptual trials, and multivariate classification
756 results in precuneus indicated the presence of both perceptual and memory-related signals. This
757 dual involvement of the precuneus in perception and memory metacognition is consistent with
758 previous studies which suggest a relationship between precuneus structure and visual perceptual
759 metacognition (Fleming et al., 2010; McCurdy et al., 2013).
760 Fourth, we found in both univariate and multivariate whole-brain analyses that domain-
761 general signals in the ventral striatum and vmPFC (including subgenual anterior cingulate
762 cortex) were modulated by confidence level. These results are compatible with previous findings
763 that have found activity in ventral striatum to be positively correlated with confidence (Daniel
764 and Pollmann, 2012; Hebart et al., 2014; Guggenmos et al., 2016). Evidence of vmPFC encoding
765 of confidence signals has been reported in connection to decision-making and value judgments
766 (De Martino et al., 2013; Lebreton et al., 2015). Our experimental design, however, does not
767 allow us to disentangle whether the signals found in these regions pertain uniquely to confidence
768 or whether they are entangled with implicit value and reward signals (for instance, the expected
769 value of being correct). Future experiments are needed to explicitly decouple reward from

770 confidence to resolve this issue.

771 In our experimental design perception and memory blocks were interleaved across runs,
772 which raises the question as to whether the domain-specific neural substrates we found would
773 persist if subjects had to switch between tasks more often. Due to inter-task "leaks" in
774 confidence (Rahnev et al., 2015), in which confidence in one task influences confidence ratings
775 on the following task (or even in the following trial), there is a possibility that interleaving
776 blocks of different tasks might favor the observation of more domain-general confidence-related
777 patterns.

778 Our experimental design assumes that visual perception and memory are distinct
779 domains. We acknowledge that distinguishing between cognitive domains or individuating
780 perceptual modalities is not straightforward (Macpherson, 2011). For instance, different
781 modalities (e.g. vision, audition, touch, etc.), different aspects within a single modality (e.g.
782 motion and color within vision), or closely related modalities (e.g. visual perception vs visual
783 short-term memory) could be part of a unified perceptual domain for metacognitive purposes.
784 Recent findings suggest metacognitive efficiency in one perceptual modality predicts
785 metacognitive efficiency in others and that they share electrophysiological markers (Faivre et al.,
786 2017). Metacognitive efficiency is also correlated across vision and visual short-term memory,
787 especially for features such as orientation (Samaha and Postle, 2017), and dACC and insula
788 regions similar to those identified here have been found to show univariate confidence signals
789 across both color and motion tasks in the visual domain (Heereman et al., 2015). However, it is
790 an open question whether more fine-grained modality-specific patterns of metacognitive activity
791 could be decoded using multivariate approaches. More research is needed on the neural
792 architecture of metacognition in other cognitive domains, and whether this architecture changes

793 in a graded or discrete fashion as a function of task or stimulus.

794 In summary, our results provide evidence for the co-existence of content-rich
795 metacognitive representations in anterior prefrontal cortex with generic confidence-related
796 signals in frontoparietal and cingulo-opercular regions. Such an architecture may be appropriate
797 for "tagging" lower-level feelings of confidence with higher-order contextual information to
798 allow effective behavioral control. Previous studies have tended to draw conclusions about either
799 domain-specific or domain-general aspects of metacognition. Here we reconcile these
800 perspectives by demonstrating that both domain-specific and domain-general signals co-exist in
801 the human brain, thus laying the groundwork for a mechanistic understanding of reflective
802 judgments of cognition.

803

804 **REFERENCES**

805

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

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

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

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

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

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

818 Botvinick MM, Cohen JD, Carter CS (2004) Conflict monitoring and anterior cingulate cortex:
819 an update. Trends Cogn Sci 8:539–546.

820 Busey TA, Tunnicliff J, Loftus GR, Loftus EF (2010) Accounts of the confidence-accuracy
821 relation in recognition memory. Psychon Bull Rev 7:26–48.

822 Chua EF, Pergolizzi D, Weintraub RR (2014) The Cognitive Neuroscience of Metamemory
823 Monitoring: Understanding Metamemory Processes, Subjective Levels Expressed, and
824 Metacognitive Accuracy. In: The Cognitive Neuroscience of Metacognition (Fleming SM,
825 Frith CD, eds), pp 267–291. Berlin: Springer.

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

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

830 Cortese A, Amano K, Koizumi A, Kawato M, Lau H (2016) Multivoxel neurofeedback
831 selectively modulates confidence without changing perceptual performance. Nat Commun
832 7:13669.

833 Daniel R, Pollmann S (2012) Striatal activations signal prediction errors on confidence in the
834 absence of external feedback. Neuroimage 59:3457–3467.

835 de Gardelle V, Mamassian P (2014) Does confidence use a common currency across two visual
836 tasks? Psychol Sci 25:1286–1288.

837 De Martino B, Fleming SM, Garrett N, Dolan RJ (2013) Confidence in value-based choice.
838 Nature Neurosci 16:105–110.

839 Donoso M, Collins AGE, Koechlin E (2014) Foundations of human reasoning in the prefrontal
840 cortex. Science 344:1481–1486.

841 Faivre N, Filevich E, Solovey G, Kuhn S, Blanke O (2017) Behavioural, modeling, and
842 electrophysiological evidence for supramodality in human metacognition. J Neurosci.

843 Flavell JH (1979) Metacognition and cognitive monitoring: A new area of cognitive–
844 developmental inquiry. Am Psychol 34:906–911.

845 Fleck MS, Daselaar SM, Dobbins IG, Cabeza R (2006) Role of Prefrontal and Anterior
846 Cingulate Regions in Decision-Making Processes Shared by Memory and Nonmemory
847 Tasks. Cereb Cortex 16:1623–1630.

848 Fleming SM (2017) Hierarchical Bayesian estimation of metacognitive efficiency from
849 confidence ratings. Neurosci Consc 3:1–14.

850 Fleming SM, Daw ND (2017) Self-evaluation of decision-making: A general Bayesian
851 framework for metacognitive computation. Psychol Rev 124:91–114.

852 Fleming SM, Dolan RJ, Frith CD (2012a) Metacognition: computation, biology and function.
853 Phil Trans R Soc B 367:1280–1286.

854 Fleming SM, Huijgen J, Dolan RJ (2012b) Prefrontal Contributions to Metacognition in
855 Perceptual Decision Making. J Neurosci 32:6117–6125.

856 Fleming SM, Lau H (2014) How to measure metacognition. Front Hum Neurosci 8:443.

857 Fleming SM, Ryu J, Golfinos JG, Blackmon KE (2014) Domain-Specific Impairment in
858 Metacognitive Accuracy Following Anterior Prefrontal Lesions. Brain 137:2811–2822.

859 Fleming SM, Weil RS, Nagy Z, Dolan RJ, Rees G (2010) Relating Introspective Accuracy to
860 Individual Differences in Brain Structure. Science 329:1541–1543.

861 Galvin SJ, Podd JV, Drga V, Whitmore J (2003) Type 2 tasks in the theory of signal
862 detectability: Discrimination between correct and incorrect decisions. Psychon Bull Rev
863 10:843–876.

864 García-Pérez MA (1998) Forced-choice staircases with fixed step sizes: asymptotic and small-
865 sample properties. Vision Res 38:1861–1881.

866 Gehring WJ, Goss B, Coles MGH, Meyer DE, Donchin E (1993) A Neural System for Error
867 Detection and Compensation. Psychol Sci 4:385–390.

868 Gorgolewski KJ, Varoquaux G, Rivera G, Schwarz Y, Ghosh SS, Maumet C, Sochat VV,
869 Nichols TE, Poldrack RA, Poline J-B, Yarkoni T, Margulies DS (2015) NeuroVault.org: a
870 web-based repository for collecting and sharing unthresholded statistical maps of the human
871 brain. Front Neuroinform 9:8.

872 Guggenmos M, Wilbertz G, Hebart MN, Sterzer P (2016) Mesolimbic confidence signals guide
873 perceptual learning in the absence of external feedback. eLife 5:e13388.

874 Hebart MN, Görgen K, Haynes JD (2015) The Decoding Toolbox (TDT): a versatile software
875 package for multivariate analyses of functional imaging data. Front Neuroinform 8:88.

876 Hebart MN, Schriever Y, Donner TH, Haynes JD (2014) The Relationship between Perceptual
877 Decision Variables and Confidence in the Human Brain. Cereb Cortex 26:118–130.

878 Heekeren HR, Marrett S, Ruff DA, Bandettini PA, Ungerleider LG (2006) Involvement of
879 human left dorsolateral prefrontal cortex in perceptual decision making is independent of
880 response modality. Proc Natl Acad Sci USA 103:10023–10028.

881 Heereman J, Walter H, Heekeren HR (2015) A task-independent neural representation of
882 subjective certainty in visual perception. Front Hum Neurosci 9:551.

883 Hilgenstock R, Weiss T, Witte OW (2014) You'd Better Think Twice: Post-Decision Perceptual
884 Confidence. Neuroimage 99:323–331.

885 Holroyd CB, Yeung N, Coles MGH, Cohen JD (2005) A Mechanism for Error Detection in
886 Speeded Response Time Tasks. J Exp Psychol Gen 134:163–191.

887 Insabato A, Pannunzi M, Rolls ET, Deco G (2010) Confidence-Related Decision Making. J
888 Neurophysiol 104:539–547.

889 Kamitani Y, Sawahata Y (2010) Spatial smoothing hurts localization but not information: Pitfalls
890 for brain mappers. Neuroimage 49:1949–1952.

891 Kelemen WL, Frost PJ, Weaver CA (2000) Individual differences in metacognition: evidence
892 against a general metacognitive ability. Mem Cognit 28:92–107.

893 Kerns JG, Cohen JD, Stenger VA, Carter CS (2004) Prefrontal cortex guides context-appropriate
894 responding during language production. Neuron 43:283–291.

895 Kiani R, Shadlen MN (2009) Representation of Confidence Associated with a Decision by
896 Neurons in the Parietal Cortex. Science 324:759–764.

897 Kruschke J (2010) Doing Bayesian Data Analysis. Academic Press.

898 Lebreton M, Abitbol R, Daunizeau J, Pessiglione M (2015) Automatic integration of confidence
899 in the brain valuation signal. Nature Neurosci 18:1159–1167.

900 Macpherson F (2011) Introduction: Individuating the Senses. In: The Senses (Macpherson F, ed),
901 pp 3–42. Oxford: Oxford University Press.

902 Maniscalco B, Lau H (2012) A signal detection theoretic approach for estimating metacognitive
903 sensitivity from confidence ratings. Conscious Cogn 21:422–430.

904 Maniscalco B, McCurdy LY, Odegaard B, Lau H (2017) Limited Cognitive Resources Explain a
905 Trade-Off between Perceptual and Metacognitive Vigilance. J Neurosci 37:1213–1224.

906 McCurdy LY, Maniscalco B, Metcalfe J, Liu KY, de Lange FP, Lau H (2013) Anatomical
907 Coupling between Distinct Metacognitive Systems for Memory and Visual Perception. J
908 Neurosci 33:1897–1906.

909 Mendoza-Halliday D, Martinez-Trujillo JC (2017) Neuronal population coding of perceived and
910 memorized visual features in the lateral prefrontal cortex. Nat Commun 8:1–13.

911 Metcalfe J, Shimamura AP eds. (1994) Metacognition. Cambridge, MA: MIT Press.

912 Nelson DL, McEvoy CL, Schreiber TA (2004) The University of South Florida free association,
913 rhyme, and word fragment norms. Behav Res Methods Instrum Comput 36:402–407.

914 Nelson TO, Narens L (1990) Metamemory: A Theoretical Framework and New Findings. In:
915 The Psychology of Learning and Motivation (Bower G, ed), pp 125–173. New York:
916 Academic Press.

917 Op de Beeck HP (2010) Against hyperacuity in brain reading: spatial smoothing does not hurt
918 multivariate fMRI analyses? Neuroimage 49:1943–1948.

919 Overgaard M, Sandberg K (2012) Kinds of access: different methods for report reveal different

920 kinds of metacognitive access. Phil Trans R Soc B 367:1287–1296.

921 Poldrack RA, Baker CI, Durnez J, Gorgolewski KJ, Matthews PM, Munafò MR, Nichols TE,
922 Poline J-B, Vul E, Yarkoni T (2017) Scanning the horizon: towards transparent and
923 reproducible neuroimaging research. Nat Rev Neurosci 18:115–126.

924 Pouget A, Drugowitsch J, Kepecs A (2016) Confidence and certainty: distinct probabilistic
925 quantities for different goals. Nature Neurosci 19:366–374.

926 Purcell BA, Kiani R (2016) Hierarchical decision processes that operate over distinct timescales
927 underlie choice and changes in strategy. Proc Natl Acad Sci USA 113:E4531–E4540.

928 R Studio Team (2015) RStudio: Integrated Development for R. Available at:
929 http://www.rstudio.com.

930 Rahnev D, Koizumi A, McCurdy LY, D'Esposito M, Lau H (2015) Confidence Leak in
931 Perceptual Decision Making. Psychol Sci 26:1664–1680.

932 Ridderinkhof KR, Ullsperger M, Crone EA, Nieuwenhuis S (2004) The role of the medial frontal
933 cortex in cognitive control. Science 306:443–447.

934 Rounis E, Maniscalco B, Rothwell JC, Passingham RE, Lau H (2010) Theta-burst transcranial
935 magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. Cogn
936 Neurosci 1:165–175.

937 Ruby E, Giles N, Lau H (2017) Finding domain-general metacognitive mechanisms requires
938 using appropriate tasks. bioRxiv:1–20.

939 Samaha J, Postle BR (2017) Correlated individual differences suggest a common mechanism
940 underlying metacognition in visual perception and visual short-term memory. Proc Royal
941 Soc B 284:20172035.

942 Scheffers MK, Coles M (2000) Performance monitoring in a confusing world: error-related brain
943 activity, judgments of response accuracy, and types of errors. J Exp Psychol Human 26:141–
944 151.

945 Schnyer DM, Verfaellie M, Alexander MP, LaFleche G, Nicholls L, Kaszniak AW (2004) A role
946 for right medial prefrontal cortex in accurate feeling-of-knowing judgments: evidence from
947 patients with lesions to frontal cortex. Neuropsychologia 42:957–966.

948 Shimamura AP, Squire LR (1986) Memory and Metamemory: A Study of the Feeling-of-
949 Knowing Phenomenon in Amnesic Patients. J Exp Psychol Learn 12:452–460.

950 Simons JS, Peers PV, Mazuz YS, Berryhill ME, Olson IR (2010) Dissociation Between Memory
951 Accuracy and Memory Confidence Following Bilateral Parietal Lesions. Cereb Cortex
952 20:479–485.

953 Sladky R, Friston KJ, Tröstl J, Cunnington R, Moser E, Windischberger C (2011) Slice-timing

954 effects and their correction in functional MRI. Neuroimage 58:588–594.

955 Tang H, Yu H-Y, Chou C-C, Crone NE, Madsen JR, Anderson WS, Kreiman G (2016) Cascade
956 of neural processing orchestrates cognitive control in human frontal cortex. eLife 5:e12352.

957 Vo VA, Li R, Kornell N, Pouget A, Cantlon JF (2014) Young Children Bet on Their Numerical
958 Skills: Metacognition in the Numerical Domain. Psychol Sci 25:1712–1721.

959 Wilson M (1988) MRC Psycholinguistic Database: Machine-usable dictionary, version 2.00.
960 Behav Res Methods Instrum Comput 20:6–10.

961 Yokoyama O, Miura N, Watanabe J, Takemoto A, Uchida S, Sugiura M, Horie K, Sato S,
962 Kawashima R, Nakamura K (2010) Right frontopolar cortex activity correlates with
963 reliability of retrospective rating of confidence in short-term recognition memory
964 performance. Neurosci Res 68:199–206.

965

966 **TABLES**

967 **Table 1. Univariate fMRI analysis - judgment-related activity interacted with domain.**
968 Significant activations at cluster-defining threshold $P<0.001$, corrected for multiple comparisons
969 at $P_{FWE}<0.05$. Conjunction of significant activations at cluster-defining threshold $P<0.001$,
970 uncorrected, of Memory (C>F) and Perception (C>F) contrasts. See Figure 3B. C=confidence,
971 F=follow
972

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 Memory (C>F) ∩ *Perception (C>F)*	Cingulate gyrus, BA32	12			6, 32, 29	R
	Insula	7			45, 11, -7	R

973

974

975 **Table 2. Univariate fMRI analysis - confidence level-related activity interacted with domain.**
976 Significant activations at cluster-defining threshold $P<0.001$, corrected for multiple comparisons
977 at $P_{FWE}<0.05$. Conjunction of significant activations at cluster-defining threshold $P<0.001$,
978 uncorrected, of Memory (M) and Perception (P) contrasts of positive and negative correlations
979 with confidence level. See Figure 3F.
980

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
	Ventral striatum	16			3, 11, 7	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

981

982 **Table 3. Judgment-related activity obtained from whole-brain searchlight classification**
983 **analyses.** Significant activations at cluster-defining threshold $P<0.001$, corrected for multiple
984 comparisons at $P_{FWE}<0.05$. See Figure 4C.
985

986

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

987

988 **Table 4. Confidence level-related activity obtained from whole-brain searchlight**
989 **classification analyses.** Accuracy maps were masked to exclude visuomotor-related activity (see
990 Figure 4D). Significant activations at cluster-defining threshold $P<0.001$, corrected for multiple
991 comparisons at $P_{FWE}<0.05$. See Figure 4F.
992

Classification	Label	Voxels at $P<0.001$	P_{FWE} cluster-corrected	Peak *z*-score	Peak voxel MNI coordinates	Laterality
Across-domain	Inferior FG, ventral striatum, 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 FG	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

993