Behavioral/Systems/Cognitive

Prefrontal Contributions to Metacognition in Perceptual Decision Making

Stephen M. Fleming,[1,2] **Josefien Huijgen,**[1] **and Raymond J. Dolan**[1]

[1]Wellcome Trust Centre for Neuroimaging, University College London, London WC1N 3BG, United Kingdom, and [2]Center for Neural Science, New York University, New York, New York 10003

Neuroscience has made considerable progress in understanding the neural substrates supporting cognitive performance in a number of domains, including memory, perception, and decision making. In contrast, how the human brain generates metacognitive awareness of task performance remains unclear. Here, we address this question by asking participants to perform perceptual decisions while providing concurrent metacognitive reports during fMRI scanning. We show that activity in right rostrolateral prefrontal cortex (rlPFC) satisfies three constraints for a role in metacognitive aspects of decision-making. Right rlPFC showed greater activity during self-report compared to a matched control condition, activity in this region correlated with reported confidence, and the strength of the relationship between activity and confidence predicted metacognitive ability across individuals. In addition, functional connectivity between right rlPFC and both contralateral PFC and visual cortex increased during metacognitive reports. We discuss these findings in a theoretical framework where rlPFC re-represents object-level decision uncertainty to facilitate metacognitive report.

Introduction

Accounting for the presence or absence of self-knowledge in simple cognitive tasks is a conundrum in cognitive science (Wilson and Dunn, 2004). In several areas of inquiry metacognition of performance is a central object of study: subjects often know when they have made an error (Rabbitt and Rodgers, 1977), appropriately scale confidence to reflect performance (Harvey, 1997), and exploit awareness of performance through advantageous wagering (Kunimoto et al., 2001; Persaud et al., 2007). However, these assessments vary in their accuracy, such that errors can be committed in the absence of awareness (Nieuwenhuis et al., 2001) and external manipulations of choice sometimes go unnoticed (Johansson et al., 2005; Logan and Crump, 2010). An influential model of metacognition postulates a dissociation between the "object" level cognition and the "meta" level, conceptualized as monitoring and controlling the object level (Nelson and Narens, 1990). Neuropsychological cases reveal dissociations between these levels (for review, see Fleming and Dolan, 2012), raising the question as to the neural mechanisms underlying metacognition in healthy individuals.

In healthy individuals, performance on a particular cognitive task and metacognition of performance are usually tightly coupled. For instance, knowing the answer to a question will tend to be accompanied by knowing that one knows the answer. Such a close relationship requires careful control over task performance to isolate a metacognitive component of behavior. Recently, by employing such a level of control, several studies have found that, during perceptual decision making, rostrolateral and dorsolateral prefrontal cortices (PFCs) are key brain areas mediating individual differences in metacognitive accuracy. The extent to which confidence tracks performance in a visual detection task is correlated both with greater gray matter volume in anterior PFC (lateral BA10) and greater white matter callosal density in tracts that connect to this region (Fleming et al., 2010b). In a related study employing functional brain imaging (fMRI), activation in right posterior-lateral BA10 correlated with metacognitive accuracy across individuals (Yokoyama et al., 2010). Furthermore, disrupting dorsolateral PFC (dlPFC) using transcranial magnetic stimulation decreases metacognitive ability without affecting task performance (Rounis et al., 2010), indicating a causal role for anterior PFC in metacognitive accuracy. However, the function of candidate prefrontal regions in metacognitive judgments of decision making is unknown.

Here we propose three criteria to triangulate brain region(s) functionally involved in metacognitive reports. First, activity should increase during metacognitive ratings compared to a matched control condition (Yokoyama et al., 2010). Second, this activity should correlate with subjective confidence. Based on previous studies of confidence in recognition memory (Henson et al., 2000; Fleck et al., 2006) and the link between rostrolateral PFC (rlPFC) recruitment and uncertainty of judgment in reasoning tasks (Christoff and Gabrielli, 2000), we expected this correlation to be negative. Third, the relationship between activity and confidence should relate to metacognitive ability (how closely confidence tracks decision performance) across individuals. Here, we implement such an approach and in so doing show that activity in right rlPFC satisfies each of these constraints.

Received Dec. 28, 2011; revised March 8, 2012; accepted March 15, 2012.

Author contributions: S.M.F. and R.J.D. designed research; S.M.F. and J.H. performed research; S.M.F. and J.H. analyzed data; S.M.F. and R.J.D. wrote the paper.

This work was supported by Wellcome Trust Programme Grant 078865/Z/05/Z to R.J.D. and a Sir Henry Wellcome Fellowship to S.M.F. The Wellcome Trust Centre for Neuroimaging is supported by core funding from the Wellcome Trust 091593/Z/10/Z.

Correspondence should be addressed to Stephen M. Fleming, Center for Neural Science, New York University, 6 Washington Place, New York, NY 10003, E-mail: fleming.sm@gmail.com.

DOI:10.1523/JNEUROSCI.6489-11.2012

 0270-6474/12/326117-09$15.00/0

Materials and Methods

Subjects. Subjects were recruited by advertisement on Gumtree and from the University College London online subject pool. Twenty-six subjects (12 males; age, 19–52 years; mean age, 28.3 years) gave written informed consent to take part in the experiment. We excluded three subjects from further analysis due to type 1 d' scores >2 SDs from the group mean. Twenty-three subjects were included in the remaining analyses (10 males; age, 19–48 years; mean age, 26.5 years). All subjects had normal or corrected-to-normal vision and no history of neurological or psychiatric illness. The study was approved by the Institute of Neurology (University College London) Research Ethics Committee.

Stimuli. Stimuli were adapted from methods and image sets documented previously (Fleming et al., 2010a). We used a set of 10 neutral faces (five male, five female) taken from the Karolinska Directed Emotional Faces set and 10 houses (photographed by S.M.F.). The stimuli were cropped to be of equal size and converted to grayscale. Fourier transforms of each image were computed, producing 20 magnitude and 20 phase matrices, and the average magnitude matrix of all stimuli was stored. On each trial, the phase matrix of a single face or house image plus a variable proportion of white noise [P(noise)] was then recombined with the average magnitude matrix, such that $P(\text{image}) = 1 - P(\text{noise})$.

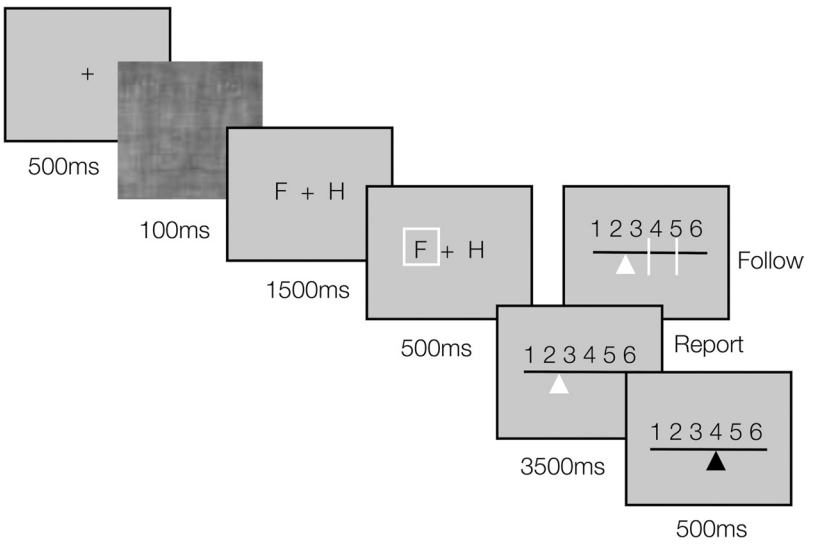

Figure 1. Perceptual decision task. On each trial, participants were asked to categorize a noisy image as either a face or a house by pressing one of two buttons held in their right hand. In the Report condition, post-decision confidence was indicated using a sliding scale. In the Follow condition, participants were instructed to slide the cursor into the zone indicated by the two vertical lines. After 3.5 s, the cursor changed color to indicate the participant's selected rating.

Face and house stimuli appeared in a random sequence, and the difficulty of each trial was controlled through use of a one-up two-down staircase procedure as used previously (Fleming et al., 2010b). After two consecutive correct responses, the amount of noise added to the image was increased by one step; after one incorrect response, the amount of noise was decreased by one step. P(noise) was adjusted in step sizes of 0.01. Independent staircases were maintained for face and house images. The aim of the staircase procedure was to equate objective performance across individuals, thus rendering metacognitive accuracy unconfounded by variation in task performance.

Task and procedure. Prior to entering the scanner, participants were familiarized with the task and confidence rating scale. Instructions emphasized that confidence ratings should reflect relative confidence, as due to the difficult nature of the task they were unlikely to be absolutely certain of their decision. During the structural scan at the start of the experiment, participants carried out 200 trials of face/house judgments (without confidence ratings) to allow the staircase parameter to stabilize. The main experiment consisted of four scanner runs, each containing 75 trials. These 75 trials were further subdivided into mini-blocks where subjects either reported their decision confidence (Report condition, 10 trials) or placed the confidence cursor in the zone indicated by the computer (Follow condition, 5 trials). Each run began with the Report condition.

Each trial consisted of the following events in sequence (see Fig. 1). A fixation cross was presented for 500 ms, followed by a face or house image for 100 ms. The participant then had 1500 ms to indicate whether the stimulus was a face or a house by pressing one of two buttons with their right hand. A square red frame appeared for 500 ms around their selected choice. No feedback was given as to whether the perceptual decision was correct or incorrect. In the Report condition, participants then indicated their decision confidence on a sliding scale, using the same two response buttons to move the cursor up or down the scale. Arbitrary scale values of 1–6 were marked on the scale at equal spacings. The confidence scale accepted participants' input for 3500 ms, followed by a change in cursor color from white to red to confirm the rating (500 ms). The final cursor position was recorded as a continuous variable on each trial. The initial cursor position on each trial was randomly jittered around the midpoint of the scale (±12% of scale length) to discourage advance motor preparation of the confidence response.

In the Follow condition, participants were asked to place the cursor in the zone indicated by the computer, specified by two vertical blue lines, instead of rating their decision confidence. The Follow condition was yoked to the Report condition such that half of the previous 10 Report ratings were randomly selected to specify the current block of Follow target positions.

After the main experiment, subjects passively viewed clear images of faces and houses as a localizer for category-specific extrastriate visual regions. As our initial whole-brain analyses did not find effects in visual cortex, we do not analyze these data further here.

Additional behavioral measures. Following scanning, participants completed the short version of the Wechsler Abbreviated Intelligence Scale (WASI; Wechsler, 1999) and the Beck Cognitive Insight Scale (BCIS; Beck et al., 2004). The short version of the WASI consists of the vocabulary and matrix reasoning subscales. Here we focus on the matrix subscale, a nonverbal measure of fluid intelligence (g) that is hypothesized to be prefrontal dependent (Roca et al., 2010), and assess the correlation between matrix score and metacognitive accuracy. One outlying subject (WASI score > 2 SD below the mean) was excluded from this analysis.

The BCIS was originally developed for use with individuals who suffer from psychotic disorders, but recent studies suggest it is a valid measure of degree of insight in healthy populations (Riggs et al., 2012). The self-reflectiveness subscale measures willingness to acknowledge fallibility, whereas the self-certainty subscale measures certainty about one's beliefs and judgments. A composite cognitive insight score is derived by subtracting the self-certainty score from the self-reflectiveness score (Riggs et al., 2012). Here we report the correlation between the composite cognitive insight score and metacognitive accuracy.

fMRI acquisition. Brain images were acquired using a 3T Allegra scanner (Siemens). BOLD-sensitive functional images were acquired using a gradient-echo EPI sequence (48 transverse slices; TR, 2.88 s; TE, 30 ms; 3 × 3 mm in-plane resolution; 2 mm slice thickness; 1 mm gap between adjacent slices; z-shim, −0.4mT/m*ms; positive phase encoding direction; slice tilt, −30°) optimized for detecting changes in the amygdala and orbitofrontal cortex (Weiskopf et al., 2006). The main experiment consisted of four runs of 202 volumes, and the localizer task consisted of a single run of 81 volumes. We also collected a T1-weighted anatomical scan and local field maps for each subject.

Behavioral data analysis. Perceptual performance was assessed using type 1 signal detection theory (SDT; Green and Swets, 1966). We calculated sensitivity (d') and bias (c) as follows:

$$d' = z(H) - z(FA)$$

$$c = -0.5 * [z(H) + z(FA)],$$

where z indicates the inverse of the cumulative normal distribution, and $H = p(\text{response} = \text{face} \mid \text{stimulus} = \text{face})$ and $FA = P(\text{response} = \text{face} \mid \text{stimulus} = \text{house})$.

Metacognitive ability (A_{roc}) was calculated as documented previously (Fleming et al., 2010b) by calculating the area under the type 2 receiver operating characteristic (ROC) using nonparametric methods (Kornbrot, 2006). Continuous confidence ratings were first binned into six quantiles. To plot the ROC, $h_i = p(\text{confidence} = i \mid \text{correct})$ and $f_i = p(\text{confidence} = i \mid \text{incorrect})$ were calculated for all i, transformed into cumulative probabilities, and plotted against each other. ROC curves were anchored at [0,0]. An ROC curve that bows sharply upward indicates that the probability of being correct rises rapidly with confidence; conversely, a flat ROC function indicates a weak link between confidence and accuracy. We note that type 2 ROC area is also affected by type 1 d' and criterion (Galvin et al., 2003; Maniscalco and Lau, 2012). Here, use of a two-alternative forced choice (2AFC) design and a continuous staircase permitted tight control over these factors, dissociating metacognitive ability from type 1 performance. The fit of the type 2 SDT model was assessed via the following linear regression model (Macmillan and Creelman, 2005):

$$z(h) = \beta_0 + \beta_1 z(f) + \varepsilon.$$

Behavioral data analysis was carried out using SPSS 17.0. Statistical tests employed are indicated at the appropriate place in Results.

fMRI preprocessing and analysis. All imaging analysis was carried out using SPM8 (Statistical Parametric Mapping; www.fil.ion.ucl.ac.uk/spm). The first five volumes of each run were discarded to allow for T1 equilibration. Functional images were realigned and unwarped using collected field maps (Andersson et al., 2001). Each participant's structural image was segmented into gray matter, white matter, and cerebral spinal fluid images using a nonlinear deformation field to map it onto template tissue probability maps (Ashburner and Friston, 2005). This mapping was applied to both structural and functional images to create spatially normalized images. Normalization is to Montreal Neurological Institute space. Normalized images were spatially smoothed using a Gaussian kernel with full-width at half-maximum of 8 mm.

fMRI time series were regressed onto a general linear model (GLM) containing stick (delta) functions representing the onset of the stimulus and boxcar functions spanning the time of the confidence rating. Separate regressors aligned to stimulus onset modeled face and house trials, each parametrically modulated by $P(\text{image})$. The confidence rating boxcar was separated into three regressors: Report following a "face" response, Report following a "house" response, and Follow; each was parametrically modulated by the selected confidence rating plus its quadratic expansion (Model 1). Regressors were convolved with a canonical hemodynamic response function (HRF). Motion correction parameters estimated from the realignment procedure were entered as covariates of no interest. Low-frequency drifts were excluded with a high-pass filter (128 s cutoff).

We specified two additional GLMs to assess alternative explanations of confidence-related activity. Model 2 separated the confidence-rating regressor into correct and incorrect trials, each parametrically modulated by reported confidence. Model 3 added type 1 reaction time (RT) as a parametric modulator of the confidence-rating regressors of Model 1, such that confidence ratings were orthogonalized with respect to reaction time.

Statistical inference. Single-subject contrast images were entered into a second-level random effects analysis using one-sample t tests against zero to assess group level significance. All reported activations survive $p < 0.05$, corrected, at the cluster level for multiple comparisons using a cluster-defining threshold of $p < 0.001$, uncorrected. Follow-up region of interest (ROI) analyses were carried out to document patterns of activation across conditions. Between-subjects correlations between brain activity and behavior were carried out by calculating the Pearson product–moment correlation between the average confidence beta on Report trials for each PFC ROI (see below, Region of interest analysis) and each individual's A_{roc}. Activations are displayed at the cluster-defining threshold of $p < 0.001$, uncorrected, using MRIcro (http://www.cabiatl.com/mricro/mricro/index.html) and Caret (Van Essen et al., 2001; http://www.nitrc.org/projects/caret/) software.

Region of interest analysis. ROIs were specified in right and left rlPFCs and right posterior parietal cortex (PPC) using the clusters shown in Figure 4*A* (negative confidence on Report trials masked by Report > Follow at $p < 0.001$, uncorrected). These clusters were transformed into binary mask images using MarsBar (http://marsbar.sourceforge.net). The dorsal anterior cingulate cortex (dACC) cluster extended into dorsal premotor cortex; thus, to maintain anatomical specificity here we defined a region of interest as a 6 mm radius sphere centered on the peak voxel in dACC [−9 17 46]. Parameter estimates and time courses averaged across all voxels in each mask were extracted using the rfxplot second-level analysis toolbox (http://rfxplot.sourceforge.net; Gläscher, 2009).

Connectivity analysis. To explore regional changes in connectivity between right rlPFC and other brain regions during metacognitive report, we carried out a psychophysiological interaction (PPI) analysis (Friston et al., 1997). PPI is a measure of context-dependent connectivity, explaining the regional activity of other brain regions in terms of the interaction between responses in a seed region (here, right rlPFC) and a cognitive or sensory process. A new design matrix was constructed that modeled the task as a block design, with separate regressors encoding Report and Follow blocks. To simplify the PPI model, we concatenated across sessions, modeling session effects with the manual entry of constant terms. We carried out PPI analysis using the generalized PPI toolbox (gPPI; http://www.nitrc.org/projects/gppi). gPPI creates a new GLM in which the deconvolved activity of the seed region is assigned to separate regressors dependent on the status of the original psychological variable (Report or Follow) and reconvolved with the hemodynamic response function. Regional time courses were extracted from the right rlPFC ROI as defined above. The main effects of Report and Follow, the seed region time course and motion parameters were included as regressors of no interest. The PPI contrast compares Report * rlPFC (+1) with Follow * rlPFC (−1). The resultant activation map reflects activity that is systematically increased in connectivity with rlPFC during Report compared to Follow trials (Fig. 5).

Results

We scanned 23 participants while they carried out a perceptual decision task followed by ratings of decision confidence (Fig. 1). We used a psychophysical staircase procedure coupled with post-decision confidence ratings to dissociate metacognitive assessments of performance—type 2 sensitivity—from performance itself, or type 1 sensitivity (Fleming et al., 2010b). We carried out three nested fMRI analyses: first, we identified brain regions increasing in activity for Report compared to Follow trials; second, we analyzed the activity of these regions for a correlation with confidence on Report trials; finally, we examined the relationship between confidence-related activity and metacognitive ability across individuals. After isolating a region of rostrolateral PFC that satisfies these three criteria, we went on to carry out an exploratory connectivity analysis seeded in rlPFC to identify candidate inter-relationships between brain regions during metacognitive reports.

Behavior

Analysis of perceptual performance showed that performance was well controlled by the staircase during scanning (mean percentage correct = 70.6 ± 3.2 SD). Importantly, type 1 d' was

similar for both Report (mean = 1.16 ± 0.25 SD) and Follow (mean = 1.14 ± 0.28 SD) trials (paired t test, $t_{(22)} = 0.23$, $p = 0.82$). This equivalence ensures that differences in brain activity between these conditions are not due to variation in perceptual task performance. Decision criteria (c) were not significantly different from zero (mean = −0.0064; one-sample t test, $t_{(22)} = -0.27$, $p = 0.79$), indicating that subjects gave a relatively equal proportion of face and house responses.

On average, participants' post-decision confidence ratings tend to increase with task performance, demonstrating good monitoring despite an absence of explicit feedback (Fig. 2*A*). To quantify individual differences in metacognitive ability, we calculated the area under the type 2 ROC (A_{roc}) using nonparametric methods (Kornbrot, 2006) as documented previously (Fleming et al., 2010b; Song et al., 2011). A_{roc} is a bias-free measure of a participant's ability to link performance to confidence. A type 2 ROC model provided an excellent fit to the confidence rating data (mean $R^2 = 0.98$). A_{roc} was independent of performance as measured by perceptual threshold ($r = -0.063$; $p = 0.78$) and Report trial type 1 d' ($r = 0.19$, $p = 0.38$), replicating previous findings of a separation between task performance and metacognitive ability (Fleming et al., 2010b; Song et al., 2011). We additionally examined whether metacognitive ability was associated with global measures of executive function and personality (see Materials and Methods). A_{roc} was positively but nonsignificantly correlated with fluid intelligence as measured by the matrix reasoning subscale of the Wechsler Abbreviated Scale of Intelligence (Wechsler, 1999) ($r = 0.19$; $p = 0.38$); prediction of A_{roc} by "cognitive insight" as measured by the Beck Cognitive Insight Scale (Beck et al., 2004) was also nonsignificant ($r = -0.19$, $p = 0.37$).

The relationship between confidence and reaction time, RT, is complicated (for review, see Pleskac and Busemeyer, 2010), but in general higher confidence is associated with faster RTs. In accordance with this prediction, RTs were negatively correlated with reported confidence (mean $r = -0.32 \pm 0.09$ SD). RTs for perceptual decisions were systematically slower on incorrect (mean = 1087 ms) compared to correct trials (mean = 971 ms; paired t test, $t_{(22)} = 10.95$, $p < 0.001$). However, the difference in RT between correct and incorrect trials was nonsignificantly related to metacognitive ability across subjects ($r = 0.26$; $p = 0.24$), rendering it unlikely that subjects directly monitored their type 1 RT to gain insight into decision accuracy. Instead, our behavioral data support an account where above-chance metacognitive ability is supported by trial-to-trial access to decision uncertainty. We next turn to our fMRI data to determine the neural basis of such a representation.

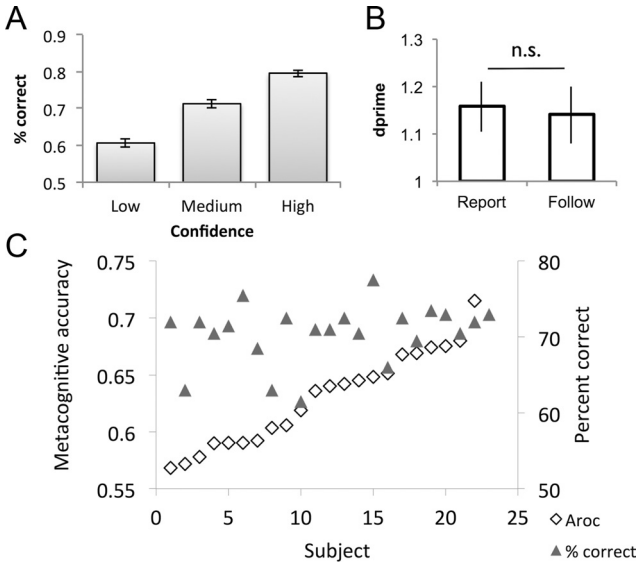

Figure 2. Behavioral results. ***A***, Performance split according to reported confidence (calculated from subject-specific tertiles). ***B***, Perceptual task performance (d') as a function of metacognition task condition. Error bars in ***A*** and ***B*** reflect the SEM. ***C***, Plot of the relationship between metacognitive accuracy (A_{roc}) and percentage correct on Report trials, with participants ordered such that A_{roc} increases from left to right.

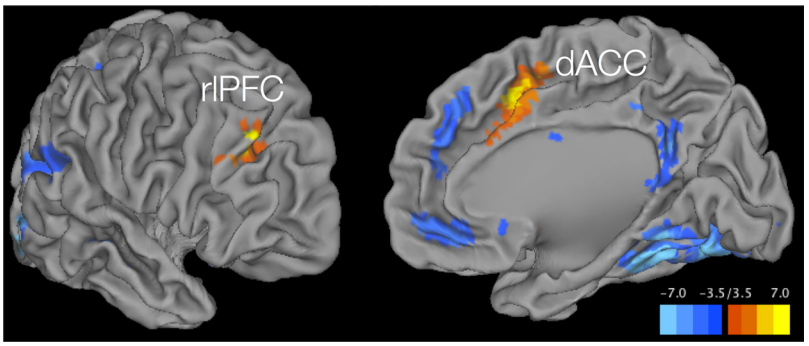

Figure 3. fMRI results. Activation in the contrasts Report > Follow (hot colors) and Follow > Report (cool colors) displayed at $p < 0.001$, uncorrected. Color bar reflects t-statistic. Significant clusters corrected for multiple comparisons at $p < 0.05$ are listed in Table 1.

Confidence-related activity

When contrasting Report > Follow, we observed increased activity in right rostrolateral PFC (rlPFC), and dorsal anterior cingulate cortex (dACC). This contrast isolates areas of increased activity when participants rated confidence in their decision compared to following the computer (Fig. 3 and Table 1). Increased activity was seen in the reverse contrast (Follow > Report) in a network of occipital and medial frontal regions (Fig. 3 and Table 1), which may reflect activation of the "task-negative" network during the less demanding control condition (Andrews-Hanna, 2012).

To identify activity satisfying our first two criteria for a role in metacognitive report (active in Report > Follow and covariation with reported confidence), we tested for correlations between activity and confidence on Report trials (using a parametric modulation analysis) masked by the previous Report > Follow contrast (using an inclusive masking threshold of $p < 0.05$, uncorrected). Negative correlations with confidence were seen in dACC, right posterior parietal cortex (R-PPC), and bilateral rlPFC (all clusters $p < 0.05$, corrected for multiple comparisons within mask volume; Fig. 4*A* and Table 2). No activity was found to correlate positively with reported confidence.

To determine the specificity of the relationship between confidence and PFC activity, we specified functional ROIs based on the correlation with confidence on Report trials (Fig. 4*B*). Follow confidence betas did not significantly differ from zero in these regions (one-sampled t tests, $t_{(22)} < 1.6$; $p > 0.1$). This result indicates that simply using the confidence scale (in the Follow condition) is not sufficient to modulate activity in R-PPC, dACC,

Table 1. Summary of significant activations for the Report > Follow and the Follow > Report contrast as reported in the main text (cluster-defining threshold, $p < 0.001$, uncorrected; all reported activations are corrected for multiple comparisons at $p < 0.05$)

Contrast	Label	Voxels at $p < 0.001$	Peak z score	p (cluster FWE corrected)	Peak voxel MNI coordinates	Laterality
Report > Follow	dACC	421	5.38	< 0.001	−3, 14, 46	L/R
	rlPFC (MFG)	73	5.10	0.040	36, 44, 28	R
Follow > Report	Occipital cortex	1894	6.79	< 0.001	−30, −88, 7	L
	Occipital cortex	1827	6.74	< 0.001	36, −85, −8	R
	Supramarginal gyrus	420	5.07	< 0.001	−57, −28, 34	L
	vmPFC	921	5.06	< 0.001	−6, 62, −11	L/R
	Posterior cingulate	561	4.57	< 0.001	0, −46, 34	L/R
	Cerebellum	104	4.71	0.010	15, −76, −41	R
	Posterior parietal cortex	77	4.28	0.033	24, −49, 64	R
	Anterior temporal lobe	121	4.15	0.005	60, −13, −14	R

FWE, Familywise error; L, left; MFG, middle frontal gyrus; MNI, Montreal Neurological Institute; R, right; vmPFC, ventromedial prefrontal cortex.

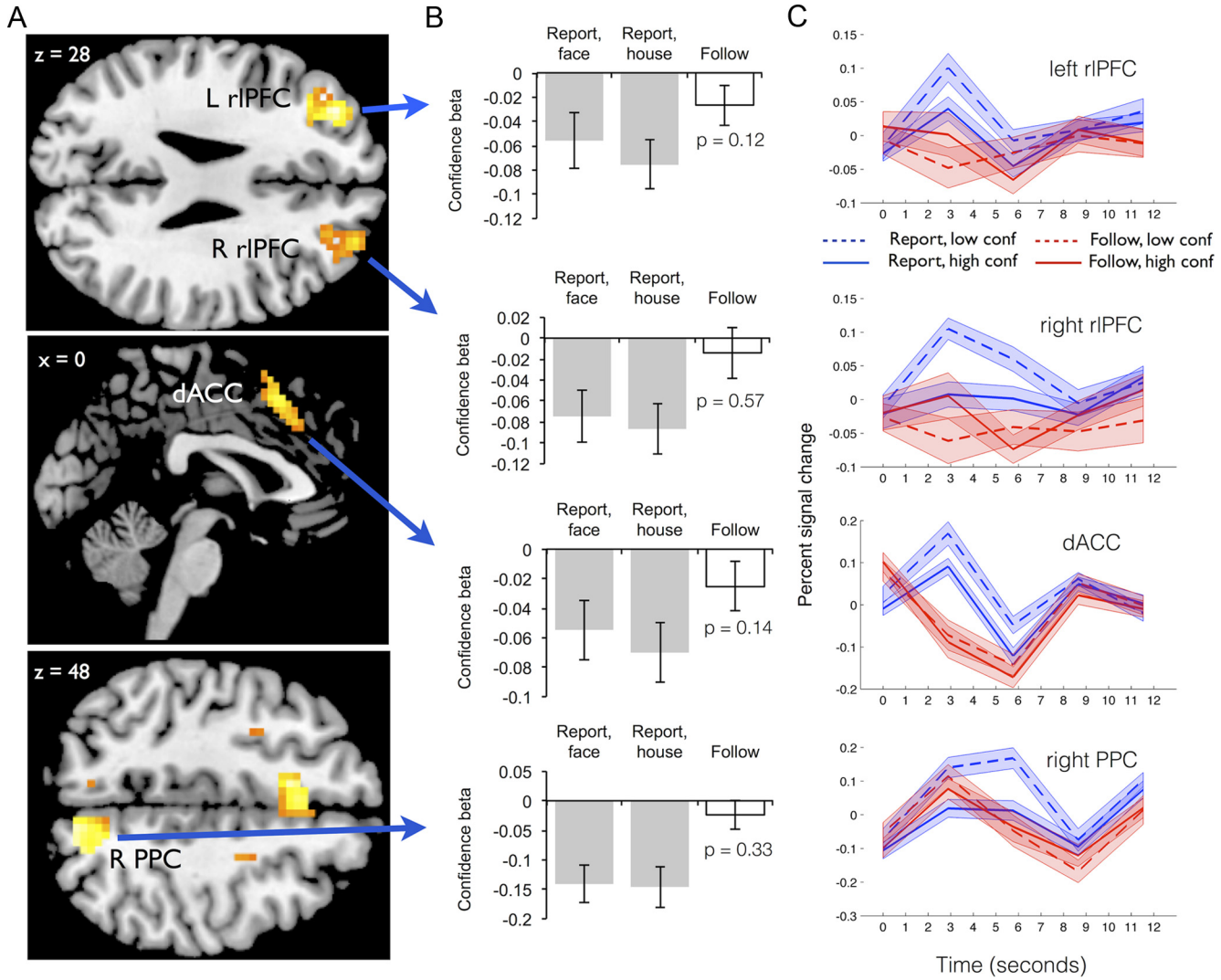

Figure 4. fMRI results. ***A***, Activation correlating negatively with confidence on Report trials, inclusively masked by Report > Follow. All clusters are significant at $p < 0.05$, corrected for multiple comparisons; images are displayed at $p < 0.001$, uncorrected. L, Left; R, right. ***B***, Parameter estimates for the correlation between confidence and activity in each ROI, split according to condition. Gray bars (Report condition) are displayed only to indicate effects of equal magnitude on face and house trials, significance having been previously established in the SPM from part ***A***. No correlation with confidence was seen in these ROIs on Follow trials (white bars; one-sample t test against zero, $p > 0.1$). Error bars reflect the SEM. ***C***, Time courses from each ROI are shown separated as a function of confidence and condition, illustrating dual effects in these regions (greater activity during Report and response to confidence during Report). Shaded areas reflect the SEM.

and rlPFC. Instead, this pattern of activity is consistent with a role for these regions in representing confidence during the Report condition. Time courses from each ROI separated by confidence and condition are displayed in Figure 4*C*, illustrating the dual effects seen in these regions (greater activity during Report and response to confidence during Report).

Finally, we tested our third criterion: prediction of metacognitive ability (A_{roc}) across subjects. For each subject, we examined

the relationship between the confidence beta on Report trials (averaged across face and house) and A_{roc} for each region of interest defined above. As the confidence beta is already negative, a stronger correlation between confidence and brain activity is expressed as a more negative relationship. Thus, we expected a negative correlation between metacognitive ability and confidence-related activity. Indeed, in right rostrolateral PFC we found a significant negative correlation between the confidence beta and A_{roc} ($r = -0.47$, $p = 0.025$; Fig. 5), which remained after controlling for Report trial type 1 d' (partial $r = -0.46$, $p = 0.032$). No significant correlation was found in left rlPFC ($r = 0.024$, $p = 0.91$). Correlations in dACC and rPPC were also negative but did not reach individual significance (dACC, $r = -0.22$, $p = 0.32$; R-PPC, $r = -0.28$, $p = 0.19$). We note that testing for regional correlations was carried out without correction for multiple comparisons, and thus a positive result here should be tempered by this caveat.

We specified two additional GLMs to assess alternative explanations of confidence-related activity in right rlPFC (see Materials and Methods). Activity did not differentiate between error and correct trials per se (Fig. 6; $t_{(22)} = 1.45$; $p = 0.16$). Furthermore, there was no significant difference between confidence betas on correct compared to incorrect trials ($t_{(22)} = 0.18$; $p = 0.86$) and confidence was significantly correlated with activity on both trial types (one-tailed t test against zero; both $p < 0.05$). Model 3 orthogonalized confidence with respect to reaction time to assess whether reaction time could account for the effect we observe in rlPFC. The effect of reported confidence remained after controlling for the variance in activity accounted for by RT (Fig. 6; $t_{(22)} = 2.35$, $p = 0.028$).

Table 2. Summary of significant activations correlating negatively with reported confidence (cluster-defining threshold $p < 0.001$, uncorrected; all reported activations are corrected for multiple comparisons at $p < 0.05$), inclusively masked by Report > Follow ($p < 0.05$, uncorrected)

Label	Voxels at $p < 0.001$	Peak z score	P (Cluster FWE corrected)	Peak voxel MNI coordinates	Laterality
dACC/pre-SMA	371	4.91	< 0.001	30, 2, 64	L/R
Posterior parietal cortex	122	4.90	0.002	15, −73, 43	R
rlPFC (MFG)	81	4.79	0.018	−33, 44, 28	L
rlPFC (MFG)	75	4.06	0.024	27, 53, 25	R

FWE, Familywise error; L, left; MFG, middle frontal gyrus; MNI, Montreal Neurological Institute; pre-SMA, pre-supplementary motor area; R, right.

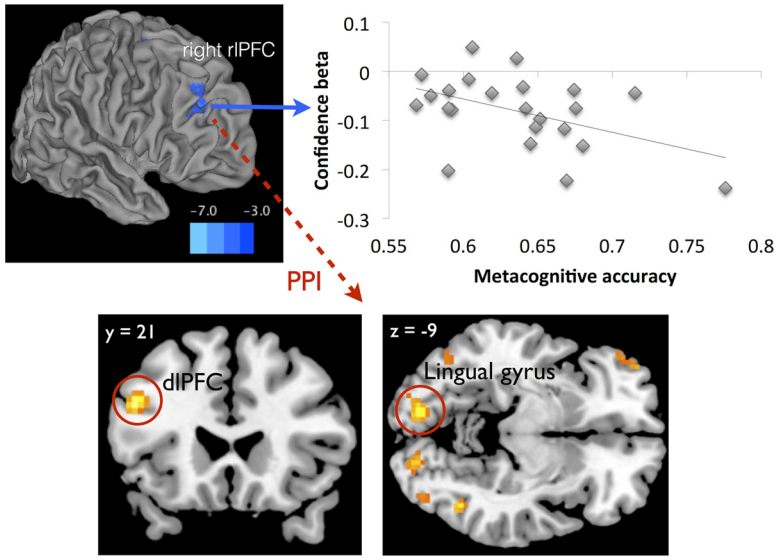

Figure 5. Individual difference and connectivity analyses. The top panel illustrates the significant correlation between confidence-related activity in right rlPFC and metacognitive accuracy across subjects. The bottom panel depicts results of an exploratory psychophysiological interaction analysis (displayed at $p < 0.001$, uncorrected), revealing whole-brain corrected ($p < 0.05$) increases in connectivity between right rlPFC and visual cortex (lingual gyrus) and between right rlPFC and left dlPFC in Report compared to Follow trials.

Functional connectivity

Our previous analyses revealed right rlPFC as satisfying our three constraints for a role in metacognitive task performance. We next explored the connectivity of this region using a psychophysiological interaction analysis (Friston et al., 1997). This analysis isolated brain regions that selectively increased in connectivity with right rlPFC during Report compared to Follow trials (Fig. 5). We found significant increases in connectivity with contralateral PFC and occipital cortex ($p < 0.05$, corrected for multiple comparisons; Table 3). Additional activations were observed in primary and extrastriate visual cortex that did not survive correction for multiple comparisons ($p < 0.001$, uncorrected; see Fig. 5). Together, this pattern of connectivity is consistent with the proposal that rlPFC integrates visual information during metacognitive Report trials.

Discussion

Here we show that fMRI activity in right rostrolateral PFC, rlPFC, satisfies three constraints for a role in metacognition of decision-making. Right rlPFC showed greater activity during self-report compared to a matched control condition, activity in rlPFC correlated with reported confidence, and the strength of the relationship between activity and confidence predicted metacognitive ability across individuals. In addition, functional connectivity between right rlPFC and both contralateral PFC and visual cortex increased during metacognitive reports.

Our work complements and extends previous studies implicating a dependence of metacognitive ability on anterior PFC structure and function (Rounis et al., 2010; Fleming et al., 2010b; Yokoyama et al., 2010). These studies identified candidate brain regions in rlPFC mediating metacognitive accuracy, but the functional role of this region in metacognition has remained unclear. By identifying a correlation between rlPFC activity and confidence that is specific to metacognitive reports, our findings may explain why damage to this region (Pannu and Kaszniak, 2005) or subclinical variation in structure (Fleming et al., 2010b) leads to altered correspondence between confidence and behavior.

Previous studies of the neural substrates of decision confidence have directly manipulated perceptual decision difficulty (Kepecs et al., 2008; Kiani and Shadlen, 2009; Rolls et al., 2010). Kiani and Shadlen (2009) showed that monkeys could choose to "opt-out" of a perceptual decision in a situation where evidence was weak, a choice predicted by firing rates of neurons in a lateral intraparietal area involved in the perceptual decision itself (Kiani and Shadlen, 2009). Similarly, firing rates in rat orbitofrontal cortex (Kepecs et al., 2008) and the fMRI signal in human ventromedial PFC (Rolls et al., 2010) encode choice difficulty as predicted by an attractor model of the decision process (Insabato

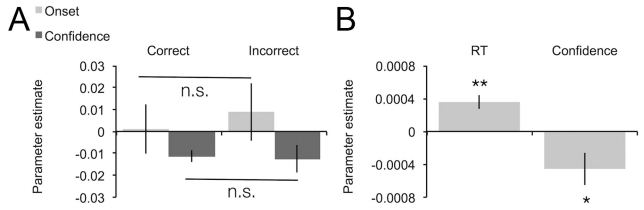

Figure 6. Analysis of right rlPFC activity using alternative GLMs (see Materials and Methods). ***A***, Model 2 separated trials into correct and incorrect to assess whether error-related activity could account for effects of confidence. Parameter estimates (arbitrary units) are shown both for average activity during confidence rating (light gray) and the confidence parametric modulator (dark gray). Neither comparison revealed significant differences between correct and incorrect trials. ***B***, Model 3 orthogonalized confidence with respect to reaction time, demonstrating that the effect of reported confidence remained after controlling for the variance in activity accounted for by RT. $^{**}p < 0.01$; $^{*}p < 0.05$. Error bars reflect the SEM.

Table 3. Summary of significant activations in a PPI examining increases in connectivity with rlPFC for Report > Follow (cluster-defining threshold, $p < 0.001$, uncorrected; all reported activations are corrected for multiple comparisons at $p < 0.05$)

Label	Peak *z* score	*p* (cluster FWE corrected)	Peak voxel MNI coordinates	Laterality
Dorsolateral PFC (MFG)	4.64	0.018	−48, 23, 28	L
Occipital cortex (lingual gyrus)	4.50	0.009	−12, −85, −8	L

FWE, family wise error; L, left; MFG, middle frontal gyrus; PFC, prefrontal cortex.

et al., 2010). These data suggest a fine-grained neural representation of choice difficulty integral to the decision process. However, in the absence of concurrent reports of decision confidence (e.g., post-decision wagers) it is difficult to assess the specificity with which this activity relates to confidence per se (Kepecs and Mainen, 2012). Instead, we suggest that these studies reveal neural representations of object-level task uncertainty that may be then re-represented for use in metacognitive report. The mechanism underlying this process, particularly across different species, remains an open question. Our finding that connectivity between right rlPFC and visual cortex is increased during metacognitive reports is indicative of such re-representation in humans.

A role in metacognition is consistent with an anatomical positioning of rlPFC at the top of a cognitive hierarchy and the fact it receives information from dorsolateral prefrontal, cingulate, and anterior temporal cortexes (Ramnani and Owen, 2004). Single-unit recordings in monkey frontopolar cortex reveal activity that tracks differences between incorrect and correct trials before receipt of feedback (Tsujimoto et al., 2010). Given that the same cells did not encode delivery of fluid reward per se, the authors concluded that monkey anterior PFC plays a role in monitoring self-generated decisions. More broadly, the contribution of rlPFC to metacognition may be to represent internal states in a format suitable for explicit communication. This hypothesis is consistent with the view that rlPFC monitors and manipulates internally generated information (Christoff and Gabrieli, 2000). Indeed, recent structural brain imaging data indicate that "reality monitoring" and metacognitive ability share a common neural substrate (Buda et al., 2011). We note however that the respective role of rlPFC (BA10) and dlPFC (BA46) in metacognitive ability is not yet clear. For example, in the present study we observed increased connectivity between right rlPFC and left dlPFC (BA46) as a function of metacognitive report. Further studies, perhaps employing single-subject analyses, are required to isolate subregions of anterior PFC critical for metacognition.

Outside of right rlPFC, we found that left rlPFC, dACC, and right PPC correlated with confidence on metacognitive report trials. Unlike in right rlPFC, confidence-related activity in these regions did not significantly predict metacognitive ability across individuals. This difference can be explained if we assume that activity in these regions correlates with object-level task uncertainty (which on average will be related to reported confidence) but does not directly support meta-level commentaries. This hypothesis is in line with reports that dACC activity increases under conflict and perceptual task difficulty (Heekeren et al., 2008; Ridderinkhof et al., 2004), while PPC is sensitive to risk in economic decision making (Huettel et al., 2005; Symmonds et al., 2011). The association between individual difference in metacognitive ability and right rlPFC is consistent with previous studies finding predominant involvement of the right side of the brain in metacognition and self-referential insight (Fleming et al., 2010b; Schnyer et al., 2004; Schmitz et al., 2006; Yokoyama et al., 2010). We note that in the present study, metacognitive ability did not significantly correlate with fluid intelligence (*g*). A recent patient study found that variance in performance on a group of tasks not accounted for by *g* was associated with lesions to right anterior PFC (Roca et al., 2010). Metacognitive ability may similarly index a capacity that is both partially independent of *g* and dependent on right PFC.

Whether rlPFC activity plays a role in metacognition in domains outside of evaluation of perceptual decision making is an open question. Control of rlPFC activity can be achieved through real-time fMRI feedback and self-regulation of attention to internal states (McCaig et al., 2011), supporting a domain-general account of its role in metacognition. Previous studies of recognition memory have shown greater activity in right dorsolateral PFC and posterior parietal cortex for low-confidence responses (Cabeza et al., 2008; Henson et al., 2000; Fleck et al., 2006), and Fleck et al. (2006) showed that confidence-related activity in lateral BA10/46 is found in both memory and decision-making paradigms. However, work examining the neural substrates of metacognition in memory also highlight the involvement of medial temporal lobe memory structures alongside PFC (Kim and Cabeza, 2007; Chua et al., 2009) and, particularly for prospective judgments, ventromedial PFC (Pannu and Kaszniak, 2005; Kao et al., 2005). Thus, further comparison of perceptual and mnemonic tasks is required to understand domain-specific and domain-general aspects of the neural substrates of metacognition.

Behaviorally, we replicate previous findings that variation in metacognitive ability can be dissociated from variation in metacognition due to task performance (Song et al., 2011; Maniscalco and Lau, 2012). This distillation of metacognition from other potentially confounding variables leads to an apparent paradox: if to measure metacognition one needs to first discount the influence of first-order behavior, then the functional benefits of metacognitive capacity would appear moot. Furthermore, adjustments in task control can be carried out in the absence of awareness of task performance (van Gaal et al., 2008). One intriguing hypothesis is that metacognitive ability instead facilitates social communication. We note a recent study of decision making found that cooperative benefit could be achieved in a perceptual decision task by sharing reports of decision confidence (Bahrami et al., 2010), and the ability to reap cooperative benefit presumably requires accurate awareness of task performance (Frith and Frith, 2012). Situating our evolving mechanistic understanding of metacognition in a functional context is an important goal for future research.

In summary, our results provide evidence that rlPFC plays a central role in linking decision confidence to metacognitive re-

ports. While confidence in a particular sensory or decision-related state may be represented in a distributed fashion across several brain regions involved in choice (Fiser et al., 2010), such representations do not explain how metacognitive reports of confidence are generated or why they may vary in their accuracy. Our findings show that the function and connectivity of rlPFC is associated with metacognitive report, potentially explaining why metacognitive deficits are observed following damage to prefrontal cortex.

References

Andersson JL, Hutton C, Ashburner J, Turner R, Friston K (2001) Modeling geometric deformations in EPI time series. Neuroimage 13:903–919.

Andrews-Hanna JR (2012) The brain's default network and its adaptive role in internal mentation. Neuroscientist. Advance online publication. Retrieved June 15, 2011. doi: 10.1177/1073858411403316.

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

Bahrami B, Olsen K, Latham PE, Roepstorff A, Rees G, Frith CD (2010) Optimally interacting minds. Science 329:1081–1085.

Beck AT, Baruch E, Balter JM, Steer RA, Warman DM (2004) A new instrument for measuring insight: the Beck Cognitive Insight Scale. Schizophr Res 68:319–329.

Buda M, Fornito A, Bergström ZM, Simons JS (2011) A specific brain structural basis for individual differences in reality monitoring. J Neurosci 31:14308–14313.

Cabeza R, Ciaramelli E, Olson IR, Moscovitch M (2008) The parietal cortex and episodic memory: an attentional account. Nat Rev Neurosci 9:613–625.

Christoff K, Gabrieli J (2000) The frontopolar cortex and human cognition: Evidence for a rostrocaudal hierarchical organization within the human prefrontal cortex. Psychobiology 28:168–186.

Chua EF, Schacter DL, Sperling RA (2009) Neural correlates of metamemory: a comparison of feeling-of-knowing and retrospective confidence judgments. J Cogn Neurosci 21:1751–1765.

Fiser J, Berkes P, Orbán G, Lengyel M (2010) Statistically optimal perception and learning: from behavior to neural representations. Trends Cogn Sci 14:119–130.

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, Dolan RJ (2012) The neural basis of metacognitive ability. Philos Trans R Soc B, in press.

Fleming SM, Whiteley L, Hulme OJ, Sahani M, Dolan RJ (2010a) Effects of category-specific costs on neural systems for perceptual decision-making. J Neurophys 103:3238–3247.

Fleming SM, Weil RS, Nagy Z, Dolan RJ, Rees G (2010b) Relating introspective accuracy to individual differences in brain structure. Science 329:1541–1543.

Friston KJ, Buechel C, Fink GR, Morris J, Rolls E, Dolan RJ (1997) Psychophysiological and modulatory interactions in neuroimaging. Neuroimage 6:218–229.

Frith CD, Frith U (2012) Mechanisms of social cognition. Annu Rev Psychol 63:287–313.

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.

Gläscher J (2009) Visualization of group inference data in functional neuroimaging. Neuroinformatics 7:73–82.

Green D, Swets J (1966) Signal detection theory and psychophysics. New York: Wiley.

Harvey N (1997) Confidence in judgment. Trends Cogn Sci 1:78–82.

Heekeren HR, Marrett S, Ungerleider LG (2008) The neural systems that mediate human perceptual decision making. Nat Rev Neurosci 9:467–479.

Henson RN, Rugg MD, Shallice T, Dolan RJ (2000) Confidence in recognition memory for words: dissociating right prefrontal roles in episodic retrieval. J Cogn Neurosci 12:913–923.

Huettel SA, Song AW, McCarthy G (2005) Decisions under uncertainty: probabilistic context influences activation of prefrontal and parietal cortices. J Neurosci 25:3304–3311.

Insabato A, Pannunzi M, Rolls ET, Deco G (2010) Confidence-related decision making. J Neurophysiol 104:539–547.

Johansson P, Hall L, Sikström S, Olsson A (2005) Failure to detect mismatches between intention and outcome in a simple decision task. Science 310:116–119.

Kao YC, Davis ES, Gabrieli JD (2005) Neural correlates of actual and predicted memory formation. Nat Neurosci 8:1776–1783.

Kepecs A, Mainen Z (2012) A computational framework for the study of confidence in humans and animals. Philos Trans R Soc B, in press.

Kepecs A, Uchida N, Zariwala HA, Mainen ZF (2008) Neural correlates, computation and behavioural impact of decision confidence. Nature 455:227–231.

Kiani R, Shadlen MN (2009) Representation of confidence associated with a decision by neurons in the parietal cortex. Science 324:759–764.

Kim H, Cabeza R (2007) Trusting our memories: dissociating the neural correlates of confidence in veridical versus illusory memories. J Neurosci 27:12190–12197.

Kornbrot DE (2006) Signal detection theory, the approach of choice: model-based and distribution-free measures and evaluation. Percept Psychophys 68:393–414.

Kunimoto C, Miller J, Pashler H (2001) Confidence and accuracy of near-threshold discrimination responses. Conscious Cogn 10:294–340.

Logan GD, Crump MJ (2010) Cognitive illusions of authorship reveal hierarchical error detection in skilled typists. Science 330:683–686.

Macmillan N, Creelman C (2005) Detection theory: a user's guide. New York: Lawrence Erlbaum.

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

McCaig RG, Dixon M, Keramatian K, Liu I, Christoff K (2011) Improved modulation of rostrolateral prefrontal cortex using real-time fMRI training and meta-cognitive awareness. Neuroimage 55:1298–1305.

Nelson TO, Narens L (1990) Metamemory: a theoretical framework and new findings. In: The psychology of learning and motivation: advances in research and theory, Vol 26 (Bower GH, ed), pp 125–173. New York: Academic.

Nieuwenhuis S, Ridderinkhof KR, Blom J, Band GP, Kok A (2001) Error-related brain potentials are differentially related to awareness of response errors: evidence from an antisaccade task. Psychophysiology 38:752–760.

Pannu JK, Kaszniak AW (2005) Metamemory experiments in neurological populations: a review. Neuropsychol Rev 15:105–130.

Persaud N, McLeod P, Cowey A (2007) Post-decision wagering objectively measures awareness. Nat Neurosci 10:257–261.

Pleskac TJ, Busemeyer JR (2010) Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. Psychol Rev 117:864–901.

Rabbitt P, Rodgers B (1977) What does a man do after he makes an error? An analysis of response programming. Q J Exp Psychol 29:727–743.

Ramnani N, Owen AM (2004) Anterior prefrontal cortex: insights into function from anatomy and neuroimaging. Nat Rev Neurosci 5:184–194.

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

Riggs SE, Grant PM, Perivoliotis D, Beck AT (2012) Assessment of cognitive insight: a qualitative review. Schizophr Bull 38:338–350.

Roca M, Parr A, Thompson R, Woolgar A, Torralva T, Antoun N, Manes F, Duncan J (2010) Executive function and fluid intelligence after frontal lobe lesions. Brain 133:234–247.

Rolls ET, Grabenhorst F, Deco G (2010) Choice, difficulty, and confidence in the brain. Neuroimage 53:694–706.

Rounis E, Maniscalco B, Rothwell JC, Passingham RE, Lau HC (2010) Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. Cogn Neurosci 1:1–11.

Schmitz TW, Rowley HA, Kawahara TN, Johnson SC (2006) Neural correlates of self-evaluative accuracy after traumatic brain injury. Neuropsychologia 44:762–773.

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.

Song C, Kanai R, Fleming SM, Weil RS, Schwarzkopf DS, Rees G (2011)

Relating inter-individual differences in metacognitive performance on different perceptual tasks. Conscious Cogn 20:1787–1792.
Symmonds M, Wright ND, Bach DR, Dolan RJ (2011) Deconstructing risk: separable encoding of variance and skewness in the brain. Neuroimage 58:1139–1149.
Tsujimoto S, Genovesio A, Wise SP (2010) Evaluating self-generated decisions in frontal pole cortex of monkeys. Nat Neurosci 13:120–126.
Van Essen DC, Drury HA, Dickson J, Harwell J, Hanlon D, Anderson CH (2001) An integrated software system for surface-based analyses of cerebral cortex. J Am Med Inform Assoc 8:443–459.
van Gaal S, Ridderinkhof KR, Fahrenfort JJ, Scholte HS, Lamme VA (2008) Frontal cortex mediates unconsciously triggered inhibitory control. J Neurosci 28:8053–8062.
Wechsler D (1999) Wechsler abbreviated scale of intelligence. San Antonio, TX: The Psychological Corporation.
Weiskopf N, Hutton C, Josephs O, Deichmann R (2006) Optimal EPI parameters for reduction of susceptibility-induced BOLD sensitivity losses: a whole-brain analysis at 3 T and 1.5 T. Neuroimage 33:493–504.
Wilson TD, Dunn EW (2004) Self-Knowledge: its limits, value, and potential for improvement. Annu Rev Psychol 55:493–518.
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.