Social Cognitive and Affective Neuroscience, 2016, 1942–1951

doi: 10.1093/scan/nsw093
Advance Access Publication Date: 21 July 2016
Original article

Neural correlates of metacognitive ability and of feeling confident: a large-scale fMRI study

Pascal Molenberghs,[1,*] Fynn-Mathis Trautwein,[2,*] Anne Böckler,[2] Tania Singer,[2] and Philipp Kanske[2]

[1]School of Psychological Sciences and Monash Institute of Cognitive and Clinical Neurosciences, Monash University, Australia and [2]Department of Social Neuroscience, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

*Joint first authors
Correspondence should be addressed to Pascal Molenberghs, Monash Institute of Cognitive and Clinical Neurosciences, 18 Innovation Walk, Clayton, Monash University, VIC, 3800, Australia. E-mail: pascal.molenberghs@monash.edu

Abstract

One important aspect of metacognition is the ability to accurately evaluate one's performance. People vary widely in their metacognitive ability and in general are too confident when evaluating their performance. This often leads to poor decision making with potentially disastrous consequences. To further our understanding of the neural underpinnings of these processes, this fMRI study investigated inter-individual differences in metacognitive ability and effects of trial-by-trial variation in subjective feelings of confidence when making metacognitive assessments. Participants ($N = 308$) evaluated their performance in a high-level social and cognitive reasoning task. The results showed that higher metacognitive accuracy was associated with a decrease in activation in the anterior medial prefrontal cortex, an area previously linked to metacognition on perception and memory. Moreover, the feeling of confidence about one's choices was associated with an increase of activation in reward, memory and motor related areas including bilateral striatum and hippocampus, while less confidence was associated with activation in areas linked with negative affect and uncertainty, including dorsomedial prefrontal and bilateral orbitofrontal cortex. This might indicate that positive affect is related to higher confidence thereby biasing metacognitive decisions towards overconfidence. In support, behavioural analyses revealed that increased confidence was associated with lower metacognitive accuracy.

Key words: metacognition; fMRI; confidence; decision making; social neuroscience

Introduction

Metacognition is the ability to think about and monitor one's own cognitive processes (Dunlosky and Metcalfe, 2009). Examples include planning a certain task, monitoring and comprehending its progress and evaluating one's own performance. In this study, we will focus on inter-individual differences in the latter aspect of metacognition, the ability to accurately evaluate one's own performance (i.e. metacognitive accuracy). A striking fact is that people are generally not very good in evaluating their own performance. The most famous example of this is the well-studied 'better-than-average' effect, the tendency for people to evaluate themselves more positively than they evaluate most other people (for reviews see Hoorens, 1993; Alicke and Govorun, 2005). For example, in one of the earliest studies documenting this effect, 94% of faculty members rated themselves as above-average teachers (Cross, 1977), while this can of course only be true for less then 50%.

The trouble with bad self-judgement and overconfidence in one's own judgment is that it can have disastrous consequences. High rates of entrepreneurial failure, global stock

Received: 22 October 2015; **Revised:** 23 June 2016; **Accepted:** 11 July 2016

market crashes, explosion of the Space Shuttle Challenger and the nuclear accident at Chernobyl have all been blamed on overconfidence (Moore and Healy, 2008). Given the importance of these problems for society in general, the aim of the current fMRI study was to get a better understanding of neural processes involved in metacognitive accuracy, the feeling of confidence in one's own judgment and the potential relation between metacognitive accuracy and overconfidence. To this end, we assessed the (accuracy of) feelings of confidence about one's own performance in a high-level factual and social reasoning task in a large sample of healthy participants (N = 308). By that, we also intended to extend previous findings about metacognition on memory and perception to higher-level cognition and to achieve sufficient statistical power, which has often been lacking in the analysis of inter-individual differences (Yarkoni *et al.*, 2009).

Previous studies have found that metacognitive accuracy can be distinguished from task performance and varies widely across individuals (for a review see Fleming and Dolan, 2012). Investigations in patients with frontal lobe lesions, for instance, showed impairments in metamemory accuracy while having preserved memory retrieval (Pannu and Kaszniak, 2005). Several lines of research point to a critical role of anterior prefrontal cortex (aPFC) structure and function as a basis of such individual differences (Fleming *et al.*, 2010; Rounis *et al.*, 2010, Yokoyama *et al.*, 2010; Baird *et al.*, 2013; McCurdy *et al.*, 2013). Furthermore, lesion studies confirmed the involvement of the aPFC in metacognition accuracy (Del Cul, *et al.*, 2009; Fleming *et al.*, 2014).

Concerning the specific association between aPFC activity and metacognitive ability, previous fMRI studies are not conclusive (some showing increase, others decrease in aPFC activity for enhanced metacognitive performance, e.g Yokoyama *et al.*, 2010, Fleming *et al.*, 2014). These fMRI studies on individual differences in metacognition accuracy used relatively small sample sizes (e.g. Fleming *et al.*, 2012: $n = 23$; Yokoyama *et al.*, 2010: $n = 25$). When performing regression analyses and working with modest effect sizes ($r = 0.2$–0.3; which is typical in psychological research) sample sizes of 100+ participants are necessary to reach 80% power (Yarkoni, 2009). If power is lower, results are less reliable, reflected in over-estimated effect-sizes or, even worse, false positives (Yarkoni, 2009). Therefore, we investigated metacognition in a large sample of more than 300 participants.

Moreover, previous studies investigating individual differences in metacognitive accuracy have focused on metacognitive judgments on simple perceptual decision making and memory performance. For example, participants had to either choose if they saw a face or a house (Fleming *et al.*, 2012), or recognize a specific dot pattern (Yokoyama *et al.*, 2010), and then indicate how confident they were in their decision. Previous studies showed that even though aPFC is involved in both metacognitive tasks, there is some specificity concerning functional connectivity and structural patterns associated with metacognitive performance in perceptual and memory tasks (Baird *et al.*, 2013, McCurdy *et al.*, 2013). What is unknown to date is what the neural correlates of metacognition on high-level cognitive processing are. We therefore applied a task that asks for confidence ratings on two types of such high-level cognitive tasks, namely inferences about others' mental states or about physical events (EmpaToM task, Kanske *et al.*, 2015).

Furthermore, previous fMRI studies have produced mixed results with regard to the neural underpinnings of subjective levels of confidence when making metacognitive assessments. For example, Chua, *et al.* (2006) found that higher subjective confidence ratings were associated with increased activation in the anterior and posterior cingulate, bilateral caudate, medial prefrontal cortex and several medial temporal lobe (MTL) regions situated around the hippocampus. Kim and Cabeza (2009) found similar MTL regions more active during high confidence trials, and additionally several prefrontal regions more active during *less* confidence trials. Moritz and colleagues (2006) found that an increase in recognition confidence was associated with the MTL and anterior and posterior cingulate cortex, while less confidence was associated with the superior posterior parietal cortex. Fleming *et al.* (2012) found no positive correlation with confidence during a face *vs* house discrimination task but less confidence was associated with increased activation in dorsal anterior cingulate cortex, right posterior parietal cortex and bilateral rostral lateral prefrontal cortex. Although it seems that the association between MTL regions and increased confidence is fairly consistent, the role of the other brain regions seems less clear-cut, especially in relation to less confidence. Again, a large sample size could provide further insights into which effects are reliable and which are not.

Finally, we explored whether overconfidence is associated with poorer metacognitive ability. Fleming *et al.*(2014) have coined the term 'metacognitive bias' (i.e. a difference in subjective confidence despite basic task performance remaining constant) and we aimed at investigating whether such a bias (overconfidence) is related to poorer metacognitive ability.

Taken together, the main goals of the this study were: (i) to investigate the neural basis of inter-individual differences in metacognitive accuracy in a large sample that allows reliable assessment; (ii) to assess metacognition not on a simple perception or memory task, but during high-level inferential cognitive processes; (iii) to assess neural networks underlying subjective confidence; and (iv) to study the relationship between overconfidence and metacognitive ability.

Methods

Participants

In total, 332 participants participated in the study, which was part of the ReSource project (Singer et al, 2015), a large-scale longitudinal study focused on investigating the effects of mental training. Twenty-four participants had to be excluded due to study dropout ($n = 5$), dropout from MRI measurements ($n = 1$) or missing data due to technical, scheduling or health issues ($n = 18$), leaving a final sample of 308 participants (age mean = 41 years, s.d. = 9, 178 female, 283 right-handed), who completed the task successfully. All participants gave written informed consent in accordance with the declaration of Helsinki. The study was approved by the Ethics committees of the University of Leipzig and the Humboldt University Berlin.

Task: EmpaToM

Metacognition was measured at the end of each trial of the EmpaToM task, which has been described in detail elsewhere (Kanske *et al.*, 2015). The task was designed to measure empathy, compassion, Theory of Mind (ToM), and metacognition. During the EmpaToM, participants are presented with a sequence of stimuli in each trial (Figure 1). After a fixation cross (1–3 s), the name of a person (1 s) who would subsequently be speaking in a short video (~15 s) was presented. Exemplary video stories and questions can be found in Supplementary

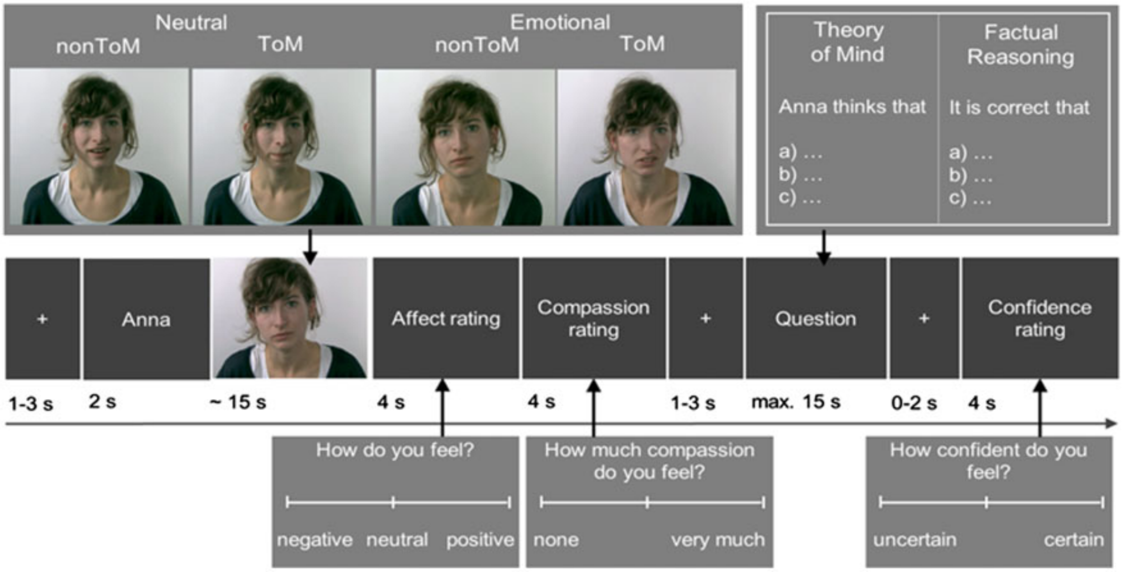

Fig. 1. EmpaToM trial sequence. Following a 2 (Emotionality of the Video) × 2 (ToM Requirements) design, four different video types were presented for each actor: Emotionally negative and neutral videos; videos with and without ToM demands, thereby leading to ToM *vs* factual reasoning questions. After each video, participants rated their own affect and their compassion for the person in the video. Subsequently, they answered a ToM or non-ToM (i.e. factual reasoning) question about the video. After each question, participants rated their confidence regarding their performance in the question. These ratings were used to calculate metacognition performance.

Material (Supplement S1). The videos differed in emotionality (emotionally neutral *vs* negative contents) and in what question they gave rise to (ToM *vs* nonToM). In total, 48 trials were presented (12 per condition). After each video, participants were asked to rate how they felt (on a scale from negative to positive; 4 s) and how much compassion they felt for the person in the previous video (scale from none to very much; 4 s). After a fixation cross (1–3 s), a multiple choice question with three response options was presented. The questions demanded either ToM reasoning or factual reasoning on the contents of the previous video. Participants had a maximum of 14 s to select one of the response options, which was then highlighted and remained on the screen for another second. After a fixation cross (0–2 s), a confidence rating was presented asking participants how confident they were to have chosen the correct response in the previous question (4 s). The confidence rating consisted of a continuous rating scale and the rated position was coded in values from 1 (uncertain) to 720 (certain). The rating scale numbers were not visible to the participant on the rating scale, but six sections indicated by seven reference lines (1–2: 1–120; 2–3: 121–240; 3–4: 241–360; 4–5: 361–480; 5–6: 481–600; 6–7: 601–720) were given to guide people's responses. In the present manuscript, we focus on the confidence ratings.

Measure of metacognitive ability: receiver operating characteristic (ROC)

To take response bias into account, signal-detection-theory (Green and Swets, 1966) was used to quantify individual differences in metacognitive ability (here defined as the ability to accurately evaluate one's own cognitive performance). Given the three-way multiple choice set-up of the metacognition for cognition task, we did not apply the meta d-prime (Maniscalco and Lau, 2012) metric, but instead chose the well-established receiver operator characteristic (ROC) to quantify meta-cognitive accuracy (Fleming and Lau, 2014). Similar to Fleming *et al.* (2010) we constructed type II ROC curves for each participant. We used the perfcurve function in Matlab (Version 8.5, Mathworks Inc., Natick, MA), entering accuracies in the task (0 or 1) as the predicted state variable (i.e. true class labels) and binned confidence ratings (1–6, corresponding to the six sections of the reference lines: 1–120; 121–240; 241–360; 361–480; 481–600; 601–720) as prediction scores. The function calculates true positive rate and false positive rate for all possible thresholds of the confidence rating. These 'hit' and 'false alarm' rates describe the points of the individual, empirical ROC curves. As a measure of overall metacognitive ability, we computed the area under the curve (AUC) using trapezoidal approximation as implemented in the perfcurve function. Higher AUC scores indicate better metacognitive ability.

If overconfidence is associated with poorer metacognitive ability, we would expect a negative correlation between the two. However in general increased confidence is associated with higher accuracy in the primary task and this was also the case in our study ($r = 0.26$; $n = 308$; $P < 0.001$). Therefore, it is important to control for mean task accuracy when measuring metacognitive bias (i.e. a difference in subjective confidence despite basic task performance remaining constant; Fleming *et al.*, 2014). To investigate the relationship between overconfidence and metacognitive ability we performed a partial correlation between confidence and the AUC score, while controlling for mean task accuracy.

MRI data acquisition

Brain images were acquired on a 3T Siemens Verio scanner (Siemens Medical Systems, Erlangen), equipped with a 32-channel head coil. Structural images were acquired using a MPRAGE T1-weighted sequence (TR = 2300 ms; TE = 2.98 ms; TI = 900; flip angle = 9°; 176 sagittal slices; matrix size = 256 × 256; FOV = 256 mm; slice thickness = 1 mm), yielding a final voxel size of 1 × 1 × 1 mm. For the functional imaging, a T2*-weighted echo-planar imaging (EPI) sequence was used (TR = 2000 ms; TE = 27 ms, flip angle = 90°). Thirty-seven axial slices were acquired covering the whole brain with a slice thickness of 3 mm, in-plane resolution 3 × 3 mm, 1 mm interslice gap,

FOV = 210 mm; matrix size 70 × 70. Each run began with three dummy volumes that were discarded from further analysis.

fMRI data analysis

Images were analysed using SPM8 (Wellcome Department of Imaging Neuroscience, London, UK). All volumes were coregistered to the SPM single-subject canonical EPI image, slice-time corrected and realigned to the mean image volume in order to correct for head motion. A high resolution anatomical image of each subject was first coregistered to the SPM single-subject canonical T1 image and then to the average functional image. The transformation matrix obtained by normalizing the anatomical image was then used to normalize functional images to MNI space. The normalized images (3 mm isotropic voxel) were spatially smoothed with a Gaussian kernel of full-width half-maximum at 8 mm. A high-pass temporal filter with cutoff of 128 s was applied to remove low-frequency drifts from the data.

After preprocessing, as part of the first level of analysis, statistical analysis was carried out using the general linear model (Friston *et al.*, 1994). Onset and duration of the four video types (neutral non-ToM; neutral ToM; emotional non-ToM; emotional ToM), their corresponding questions and the rating periods (affect, compassion and confidence) were modeled. These regressors were convolved with a canonical hemodynamic response function (HRF). Effects of head motion were accounted for by modeling the six motion parameters for each subject as effects of no interest in the design matrix. To further reduce influence of potential noise-artifacts, we used the RobustWLS Toolbox (Diedrichsen and Shadmehr, 2005), which down-weights images with higher noise variance through a weighted-least-squares approach.

To measure neural correlates of inter-individual differences in metacognitive accuracy, a 'metacognition contrast' was created in the first-level analysis by subtracting the average BOLD response of the affect and compassion ratings from the confidence rating for each participant. This contrast was related to behavioural metacognitive accuracy in the second-level analysis (see below). To measure effects of trial-by-trial variation of subjective feelings of confidence, the rated confidence (1–720) for each trial was included as a parametric modulation of the confidence rating period (i.e. the time during which people rate how confident they were about their decision). To make sure that the effects of rated confidence were not confounded by motor processes involved in making the rating, the duration of the rating process (time till last button press) and moved distance of the rating marker (distance between start and end marker position) were additionally included in the design as parametric modulators of the confidence rating regressor. Since these were entered into the design matrix prior to the rated confidence regressor, the sequential orthogonalization implemented in SPM would remove variance related to the duration and extent of button presses from the rated confidence. Furthermore, we also checked whether the effect of confidence on neural activity would differ across the four conditions of the 2 × 2 factorial task design [emotionality (neutral *vs* emotional) by ToM demand (non-ToM *vs* ToM question)]. To this end, we ran additional models with separate regressors for confidence rating epochs and their parametric modulation for the four conditions (i.e. neutral non-ToM, neutral ToM, emotional non-ToM, emotional ToM).

In the second level of analysis, first-level contrast images for the 'metacognition contrast' [confidence rating−(affect + compassion rating)/2] were entered into an SPM regression analysis with the AUC score (i.e. inter-individual differences in metacognitive accuracy) as the only regressor. This analysis was repeated for AUC scores calculated separately on neutral, emotional, non-ToM and ToM trials in order to check for consistency across these conditions. To assess intra-individual differences in subjective feelings of confidence, first-level parameter estimates of subjective confidence ratings were entered into a one-sample *t*-test for random effects analysis at the second-level. Additionally, differences and commonalities across the task conditions were checked by entering contrasts of the four parametric modulators into a 2 × 2 factorial design (emotionality × ToM demands in the question) and evaluating effects separately for each of the two factor levels as well as contrasting them against each other. Significant activity for all analyses was defined by a voxel-wise FWE of $P < 0.05$ corrected for the whole brain.

Results

Behavioural results

Participants' mean AUC score was 0.69 (range: 0.46–0.92; s.d. = 0.094), mean confidence score was 479 (range: 210–696; s.d. = 84) and mean accuracy was 61% (range: 31–88%; s.d. = 12%; split per condition: neutral non-ToM = 56.9%; neutral ToM = 63.8%; emotional non-ToM = 55.3%; emotional ToM = 66.8%). A significant negative partial correlation was found between the confidence and AUC score ($r = -0.126$; $n = 308$; $P = 0.027$), when controlling for mean accuracy.

It is possible that the negative correlation was a spurious result of how we measured (i.e. using trapezoidal approximation) AUC. The main computational issue with trapezoidal approximation is that it systematically underestimates area under the 'true' ROC curve composed of infinitely many points, and this underestimation can become more pronounced as response bias increases (Hanley and McNeil, 1982). Thus, this issue of underestimation could possibly explain why confidence and AUC are negatively correlated. This potential confound is explained in detail in Supplementary Figure 1.

To rule out the potential deflationary account, we also estimated the ROC curve using the parametric binormal ROC model (with matlab code provided by Brown and Davis, 2006). This method does not rely on trapezoidal approximation and thus should be robust to the bias of underestimation. However, a disadvantage of this method is that it relies on the assumption that the prediction scores of the two classes (i.e. correct and incorrect trials) follow a binormal distribution or can be monotonously transformed to achieve a normal distribution, an assumption that is difficult to assess in practice. Therefore, we used the classical nonparametric curves for all analyses in the manuscript and only checked the critical partial correlation using the binormal ROC model. AUC scores for both methods were highly correlated ($r = 0.98$) and had very similar mean values (0.690 and 0.696), arguing for the validity of the parametric approach. Importantly, running the partial correlation analysis on the new AUC scores showed a similar result ($r = -0.11$, $p = 0.06$).

fMRI results

Inter-individual difference in metacognitive ability: AUC score. A significant negative correlation between the AUC score and the metacognition contrast was found in the anterior medial prefrontal cortex (aPFC; −9, 54, 15; T = 4.68; extent = 3 voxels;

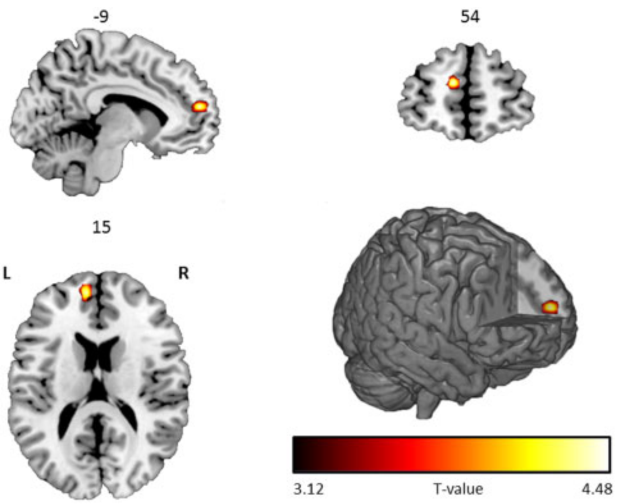

Fig. 2. Significant negative correlation between the AUC score and the metacognition contrast, shown on a rendered image and coronal, axial and sagital transversal slices with numbers above each slice representing coordinates in MNI space. L corresponds to left and R to the right side of the brain. Activations are displayed on a ch256 template using MRIcroGL (http://www.mccauslandcenter.sc.edu/mricrogl/) software.

P FWE = 0.025; Figure 2). The reverse (positive) correlation revealed no significant effects. Since inter-individual differences in primary task performance (i.e. accuracy in factual and mental state reasoning) were correlated with AUC scores ($r_{\text{overall accuracy, AUC}} = 0.37$, $P < 0.001$; $r_{\text{ToM accuracy, ToM AUC}} = 0.11$, $P = 0.06$; $r_{\text{non-ToM accuracy, non-ToM AUC}} = 0.32$, $P < 0.001$; $r_{\text{emotional accuracy, emotional AUC}} = 0.21$, $P < 0.001$; $r_{\text{neutral accuracy, neutral AUC}} = 0.25$, $P < 0.001$), an additional analysis was performed to make sure that the negative correlation between the AUC score and the metacognition contrast in the aPFC was not confounded by accuracy on the primary task. Performance (i.e. mean accuracy) in the main task was regressed out of the total AUC score and the same analysis was performed again in SPM using an ROI approach (the aPFC region from our previous analysis as displayed in Figure 2 was used as an ROI). The analysis (thresholded at $P = 0.05$ uncorrected and FWE corrected for the size of the ROI) revealed a very similar negative correlation between the AUC score and the metacognition contrast in aPFC (−9, 54, 18; $T = 4.18$; P FWE < 0.001), showing that our main result was not confounded by inter-individual variability in performance on the primary task itself.

To check the consistency of the negative correlation between the AUC score and the metacognition contrast in the aPFC, we ran the same analysis for AUC scores calculated separately for ToM, non-ToM, emotional and neutral conditions using an ROI-based approach (the region displayed in Figure 2 was used as an ROI). The results (thresholded at $P = 0.05$ uncorrected and FWE corrected for the size of the ROI) were similar for each of the four analyses (Non-ToM: −9, 54, 15; $T = 4.82$; P FWE < 0.001; ToM: −12, 57, 15; $T = 2.44$; P FWE = 0.06; Emotional: 12, 57, 15; $T = 3.36$; P FWE = 0.006; Non-Emotional: −9, 54, 18; $T = 4.04$; P FWE = 0.001). No significant correlation between the condition specific AUC scores was observed outside aPFC in whole brain analyses. Furthermore, contrast analyses between these regressors (ToM *vs* non-Tom and Emotional *vs* Non-Emotional) revealed no significant differences between the conditions. This pattern shows that the result was highly consistent across the four conditions and thus relates to metacognition on higher-level reasoning in general, and not to the specific contents of any individual condition.

Intra-individual differences in confidence ratings. An association with *less* confidence was found in the dorsal and anterior medial prefrontal cortex (extending into neighbouring dorsal anterior cingulate and supplementary motor area) and bilateral orbitofrontal gyrus (extending into neighbouring anterior insula and inferior frontal gyrus). Other regions significantly associated with less confidence were the left lingual gyrus, right superior occipital gyrus (extending into superior parietal lobe), right supramarginal gyrus and left cerebellum. For a full description of the results see Table 1 and Figure 3.

Brain regions associated with *more* confidence were found in bilateral striatum (including caudate and putamen), bilateral hippocampus and postcentral gyrus (extending into neighbouring precentral gyrus and supplementary motor area). Other regions significantly associated with more confidence were the right lingual gyrus extending into the bilateral cuneus, left lingual gyrus and right anterior and posterior middle temporal gyrus. For a full description of the results Table 1 and Figure 3.

We also analysed confidence-related activations individually for the four conditions of the 2 × 2 factorial design [emotionality (neutral *vs* emotional) by ToM demand (non-ToM *vs* ToM question). For this, we ran additional first-level models with separate regressors for confidence rating epochs and their parametric modulators of the four conditions. Contrasts of the four parametric modulators were then entered into a 2 × 2 factorial design (emotionality × ToM demands in the question). This analysis yielded similar activations for each of the conditions (Supplementary Figures 2 to 5). Contrasting emotional *vs* neutral and ToM *vs* non-ToM showed no significant activations. A more sensitive ROI approach (small volume correction within a mask defined by the overall confidence analysis) yielded stronger confidence-related activations in striate regions for ToM compared to non-ToM trials (Supplementary Figure 6), while there were no significant differences between neutral and emotional trials.

Discussion

Incorrectly evaluating one's own performance (i.e. poor metacognitive ability), and especially overconfidence, have been suggested to be common causes of faulty decision-making with potentially serious consequences (Moore and Healy, 2008). In order to better understand the involved processes we intended to (i) assess neural correlates of inter-individual differences in metacognitive accuracy in a large sample that allows reliable assessment; (ii) assess metacognition during high-level inferential cognitive processes (iii) assess neural networks underlying subjective confidence (iv) study the relationship between overconfidence and metacognitive ability.

Neural correlates of inter-individual differences in metacognitive accuracy

Increased metacognitive accuracy across individuals was associated with less activation in a medial part of aPFC when making metacognitive assessments. This result corresponds well with previous studies linking structural and functional properties of aPFC with metacognitive ability (Baird *et al.*, 2013; Del Cul, *et al.*, 2009; Fleming *et al.*, 2010; Rounis *et al.*, 2010; Yokoyama *et al.*, 2010; Fleming *et al.*, 2012; McCurdy *et al.*, 2013; Fleming *et al.*, 2014).

Table 1. Cluster size and associated peak values for the significant brain regions associated with feeling less and more confident during metacognition assessment. Only clusters with more than 10 voxels are reported.

	Cluster size	Peak *P* value	Peak *T* value	MNI coordinates		
				x	y	z
Less confident						
Left lingual gyrus	209	<0.001	14.92	−9	−75	−6
Dorsal superior frontal gyrus	2288	<0.001	12.10	12	15	60
Midcingulate gyrus		<0.001	10.04	−3	24	39
Right anterior superior frontal gyrus		<0.001	9.87	24	54	27
Right superior occipital gyrus	328	<0.001	9.58	42	−81	27
Right superior parietal lobe		<0.001	9.17	15	−63	54
Left anterior superior frontal gyrus	256	<0.001	9.38	−24	51	21
Left anterior insula	236	<0.001	8.41	−33	18	−12
Left inferior frontal gyrus		<0.001	7.03	−45	18	3
Left orbitofrontal gyrus		<0.001	5.01	−51	36	−9
Right orbitofrontal gyrus	291	<0.001	8.28	51	24	−3
Right orbitofrontal gyrus		<0.001	6.10	45	39	−9
Right orbitofrontal gyrus		<0.001	6.03	33	21	−12
Right supramarginal gyrus	30	<0.001	6.36	60	−48	30
Left cerebellum	17	0.001	5.69	−33	−60	−30
More confident						
Right lingual gyrus	1937	<0.0001	20.31	12	−66	−3
Right cuneus		<0.0001	11.04	15	−81	33
Left cuneus		<0.0001	9.97	−6	−90	18
Right Caudate	7715	<0.0001	12.96	12	12	−6
Left Caudate		<0.0001	12.57	−9	12	−3
Left Postcentral gyrus		<0.0001	12.10	−42	−27	60
Left lingual gyrus	52	<0.0001	7.94	−9	−90	−12
Left lingual gyrus		<0.0001	5.90	−21	−87	−9
Right anterior middle temporal gyrus	65	<0.0001	6.02	60	−3	−12
Right posterior middle temporal gyrus	18	0.005	5.19	48	−66	9

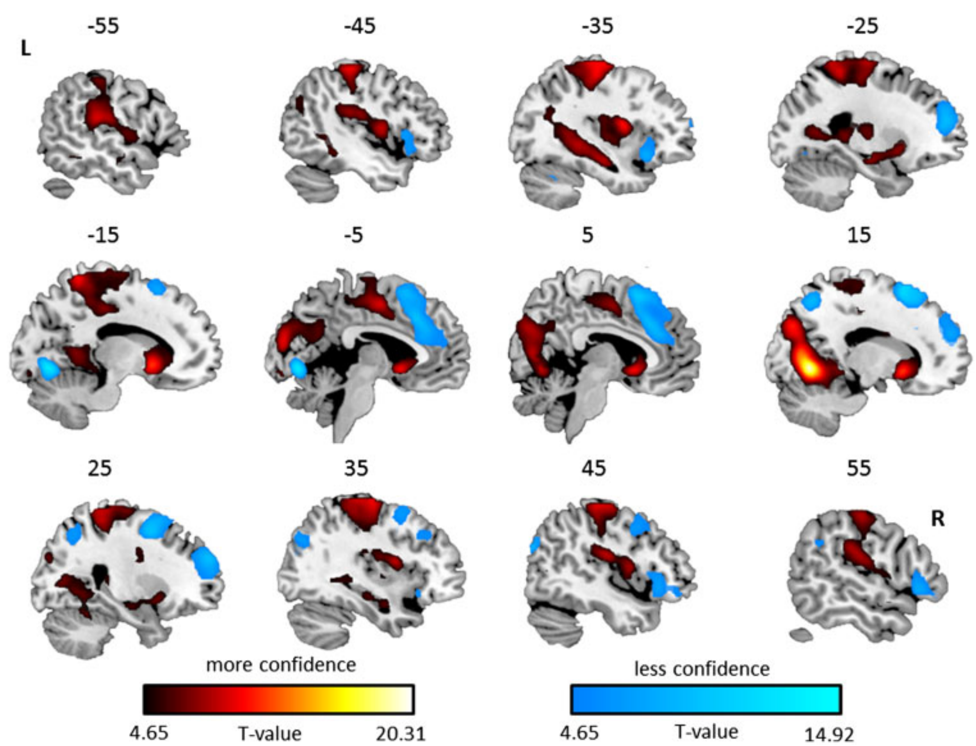

Fig. 3. Significant brain regions associated with *more* (red) and *less* (blue) confidence, shown on sagittal slices with numbers above each slice representing coordinates in X MNI space. L corresponds to left and R to the right side of the brain. Activations are displayed on a ch256 template using MRIcroGL (http://www.mccauslandcenter.sc.edu/mricrogl/) software.

In contrast to Yokoyama *et al.* (2010) who reported a positive correlation between metacognitive accuracy and brain activation in aPFC, the correlation was negative in our study. Another fMRI study which tested correlations between inter-individual differences in metacognition and brain function also found a negative effect in aPFC (Fleming *et al.*, 2012). However, this result is more difficult to interpret in terms of directionality, as 'brain function' here refers to the within subject strength of association between confidence and aPFC activity, which was in itself negative. More broadly, most studies that have investigated relationships between cognitive abilities and brain activity have found negative effects, especially in frontal brain areas, favouring a neural efficiency hypothesis (for reviews see Grill-Spector *et al.*, 2006; Neubauer and Fink 2009). Moreover, performance impairment is also typically associated with a reduction in neural efficiency (e.g. Kanske *et al.*, 2013; Wessa *et al.*, 2013). However, there is also evidence that contextual factors such as (subjective) task difficulty can induce positive correlations (Neubauer and Fink, 2009), potentially explaining the divergent finding from Yokoyama *et al.* (2010). Thus, the fact that we find a negative correlation between metacognitive accuracy and brain activation in aPFC is not entirely surprising.

Another aspect that has to be taken into account when testing brain-behaviour relationships in terms of inter-individual differences is sufficiency of statistical power and multiple comparison correction (Yarkoni, 2009). As Fleming *et al.* (2010) noted, 'testing for regional correlations was carried out without correction for multiple comparisons, and thus a positive result here should be tempered by this caveat.' Moreover, the effect sizes expected for such analyses of individual differences require large sample sizes in order to avoid inflation of effects or Type II errors due to low statistical power (Yarkoni, 2009). Therefore, it was important to demonstrate that inter-individual differences in aPFC activation are associated with differences in metacognitive accuracy in a study that has sufficient power and allows for multiple comparison correction, as done in the present study.

Metacognition in the context of high-level inferential cognitive processes

Previous studies on the neural processes involved in metacognition focused on metacognitive judgements of one's own performance in memory and perception tasks (Del Cul, *et al.*, 2009; Fleming *et al.*, 2010; Rounis *et al.*, 2010; Yokoyama *et al.*, 2010; Fleming *et al.*, 2012; Baird *et al.*, 2013; McCurdy *et al.*, 2013; Fleming *et al.*, 2014). Interestingly, it seems that similar to those studies, inter-individual differences in metacognition on high-level reasoning also relate to aPFC function. This suggests that this region is involved in metacognitive assessments in different domains. This was further confirmed by analyses that differentiated between different types of questions of the primary task (mental state inference *vs* factual reasoning and neutral *vs* emotional context of the question), which found similar associations with aPFC activity in each of these conditions.

The role of the aPFC in metacognition is also in line with a review of neuroimaging studies by Christoff and Gabrieli (2000) who found that the aPFC is often involved when internally generated information needs to be evaluated. Other evidence from non-human primate studies further point to a critical role of the aPFC in metacognition. For example, single-cell recordings in monkeys revealed that specific neurons in the aPFC do not respond when a monkey makes a decision, but respond later on when the monkey evaluates its own decision (Tsujimoto *et al.*, 2010). Humans have better metacognitive abilities than non-human primates and it has been suggested that the relative larger size (compared to the rest of the brain) of the aPFC in humans versus apes is the reason for this (Semendeferi et al., 2001).

Beyond the general implication of aPFC in metacognition, several authors have suggested a lateral-medial separation within aPFC (Fleming and Dolan, 2012; Baird *et al.*, 2013). The peak voxel found in the present study was located more medial than in studies that investigated individual differences in metacognition on perception (Fleming *et al.*, 2010; Fleming *et al.*, 2012; Baird *et al.*, 2013) and closer to a seed region whose connectivity was correlated with metacognition on memory (Baird *et al.*, 2013). This indicates that the processes involved in metacognition on high-level reasoning are more similar to metacognition on memory than to metacognition on perception. This is consistent with the requirements of the employed task, which demands inferential processing on information seen in the previous video and presumably stored in working memory.

Trial-by-trial variation in confidence

Less confidence led to more activation in dorsal medial prefrontal cortex, bilateral anterior prefrontal cortex and midcingulate cortex. These areas correspond well with the fMRI study on metacognition by Fleming and colleagues (2012), who found that these brain areas were more active during metacognition and less confidence. Others have also associated these regions with uncertainty in decision making (Volz *et al.*, 2005; Krain *et al.*, 2006; Potvin *et al.*, 2014). In addition, we found that less confidence was also associated with more activation in bilateral lateral orbitofrontal cortex (OFC). Meta-analyses on the OFC have found that the lateral part of the OFC is typically associated with feelings of displeasure (Kringelbach and Rolls, 2004; Berridge and Kringelbach, 2013). Hsu *et al.* (2005) also showed that making decisions in uncertain situations leads to more activation in the lateral OFC and that patients with lesions in these areas are insensitive to the level of uncertainty when making choices.

More confidence was associated with more activation in bilateral striatum. The peak activation in this area was located in the caudate, an area often associated with dopamine and stimulus-action-reward association (Haruno and Kawato, 2006). Together with the increased activation seen in brain areas often involved in action execution and preparation such as the postcentral gyrus, precentral gyrus and supplementary motor area, this pattern suggests that more confidence leads to stronger connection between the motor response and the positive feelings associated with being more confident about one's own choices. Note that these results cannot be explained by duration of the (motor) response *per se* as these effects were controlled for in our fMRI analysis. The link between more confidence and the dopamine network in the striatum suggests that being confident about one's choices may be an intrinsically positive experience. Together with the evidence that brain areas associated with less confidence are typically involved in displeasure and uncertainty, this might partially explain why people in general prefer to be overconfident about their choices and intrinsic abilities (Hoorens, 1993; Alicke and Govorun, 2005). The link between overconfidence and bad judgement was confirmed in our behavioural results, which showed that when controlling for mean accuracy, confidence was negatively correlated with metacognitive ability.

More confidence was also associated with increased activation in the bilateral hippocampus. This latter result fits well

with previous metacognition fMRI studies that found this region to be associated with more confidence, and is consistent with the view that increased activation in the hippocampus is associated with better memory retrieval (Chua *et al.*, 2006; Moritz *et al.*, 2006; Kim and Cabeza, 2009). There were additional brain activations, which might be a result of the task design. For example, less confidence was related to more activation in left occipital cortex, while the opposite (more activation in right occipital cortex) was true for more confidence. This was not surprising because people systematically had to move the confidence scale to the left (leading to more visual stimulation (i.e. more letters) coming from the right visual field because their eyes have now moved to the left side of the centre) to indicate less confidence. This increased visual stimulation from the right visual field led to more activation in the left occipital cortex. The opposite (i.e. more right occipital cortex activation) is true for sliding the cursor to the right (leading to more visual stimulation from the left visual field), which indicated more confidence.

The relationship between overconfidence and metacognitive ability

The potential implication of the association between rated confidence and neural networks involved in processing positive and negative affect—namely that people tend to overestimate their confidence in order to feel more positive—has a testable implication for behavioural performance. Specifically, when controlling for the true task performance level, a stronger inclination to overestimate one's own performance could undermine metacognitive accuracy. In fact, we found that reported confidence correlated positively with metacognitive accuracy (as implied by the above chance metacognitive accuracy, cf. Fleming *et al.*, 2014). Importantly however, this correlation was negative when controlling for the subjects' actual task accuracy level. This indicates that there is a bias to overestimate one's performance, which varies between subjects and can undermine the accuracy of one's metacognitive judgement.

Note that the method used to assess metacognitive ability (trapezoidal approximation of the area under ROC curve) underestimates metacognitive accuracy, potentially to a larger extent in the presence of response bias (i.e. overconfidence). Therefore, we repeated the analysis with metacognition scores derived from an independent estimation method not susceptible to underestimation, which yielded results consistent with the first analysis. This analysis suggests that the negative correlation is a true effect rather than a spurious result because of how the area under the curve was estimated.

Strengths, limitations and future work

One important advantage of our study compared to previous fMRI studies on metacognition was the large sample size ($N = 308$). This allowed us to use stringent whole brain FWE correction and produce reliable results. Especially when using regression analysis in fMRI studies, it is critical that large sample sizes are used in order to obtain reliable results that represent true effect sizes (Yarkoni *et al*, 2009). The low power in neuroscience studies in general has been suggested as the key factor in producing results that are often not reproducible, which has led some to conclude that neuroscience research is unreliable and wasteful (Button *et al.*, 2013).

The design of the EmpaToM task allowed us to study inter-individual differences in both metacognitive ability and intra-individual differences in subjective confidence on a trial-by-trial basis. Moreover, this is the first study to probe neural correlates of metacognition on high-level social and factual reasoning. However, a potential limitation in this study was the absence of a direct control condition for the metacognitive assessment (as there was no condition during which people had to move the confidence cursor unrelated to their subjective confidence rating). Therefore, we could not identify the network for metacognitive assessment *per se*.

Future studies should further investigate if metacognitive ability relies on similar brain regions across a wide variety of tasks. The fact that we also find a similar association between the aPFC and metacognitive ability in our higher-level metacognition study as others previously did in lower level metacognitive studies (Fleming *et al.*, 2010, 2012; Yokoyama *et al.*, 2010), suggests that the aPFC might be involved in a wide variety of metacognition assessments. However, some other evidence points to the contrary. For example, McCurdy and colleagues (2013) found that although visual and memory metacognitive efficiency correlated across different tasks, visual metacognitive efficiency was correlated with brain volume in aPFC, while memory metacognitive efficiency was correlated with brain volume in the precuneus. Therefore, future metacognitive fMRI studies should try to combine different types of metacognitive assessment in a single study. This will allow more precise conclusions about how the underlying neural architecture of metacognitive ability is influenced by differences in task components and complexity. Finally, it seems worthwhile for future research to consider affective processes as a biasing factor in metacognitive self-evaluation. Future studies could explicitly relate such biases to state and trait measures of positive and negative affect. This might help to further elucidate mechanisms which lead to unrealistically high self-assessments (Alicke and Govorun, 2005), and potentially severe faulty decision making.

Funding

T.S. as principal investigator, received funding for the ReSource Project from the European Research Council under the European Community's Seventh Framework Program (FP7/20072013/ ERC grant agreement no. 205557) and from the Max Planck Society. P.M. was supported by an Australian Research Council (ARC) Early Career Research Award (DE130100120), Heart Foundation Future Leader Fellowship (100458), and an ARC Discovery Grant (DP130100559).

Acknowledgements

We are thankful to the Department of Social Neurosciences for their support with the ReSource project. In particular we want to thank Hilmar Bromer, Josefine Drößler, Johannes Mahr, Ulrike Nemeth, Lisa Nix, Lilia Papst, Sophie Pauligk for help with task development, and Manuela Hofmann, Sylvia Neubert, Nicole Pampus for help with data acquisition.

Supplementary data

Supplementary data are available at *SCAN* online.

Conflict of interest. *None declared*.

References

Alicke, M.D., Govorun, O. (2005). The better-than-average effect. In Alicke, M.D., Dunning D., Krueger J., editors. *The Self in Social Judgment* (pp. 85–106). New York: Psychology Press.

Baird, B., Smallwood, J., Gorgolewski, K.J., Margulies, D.S. (2013). Medial and lateral networks in anterior prefrontal cortex support metacognitive ability for memory and perception. *The Journal of Neuroscience*, **33**(42), 16657–65.

Berridge, K.C., Kringelbach, M.L. (2013). Neuroscience of affect: brain mechanisms of pleasure and displeasure. *Current Opinion in Neurobiology*, **23**(3), 294–303.

Brown C.D., Davis H.T. (2006). Receiver operating characteristics curves and related decision measures: a tutorial. *Chemometrics and Intelligent Laboratory Systems*, **80**(1), 24–38.

Button, K.S., Ioannidis, J.P., Mokrysz, C., *et al.* (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, **14**(5), 365–76.

Christoff, K., Gabrieli, J.D. (2000). The frontopolar cortex and human cognition: Evidence for a rostrocaudal hierarchical organization within the human prefrontal cortex. *Psychobiology*, **28**(2), 168–86.

Chua, E.F., Schacter, D.L., Rand-Giovannetti, E., Sperling, R.A. (2006). Understanding metamemory: neural correlates of the cognitive process and subjective level of confidence in recognition memory. *Neuroimage*, **29**(4), 1150–60.

Cross, P. (1977). Not can but will college teachers be improved? *New Directions for Higher Education*, (17), 1–15.

Del Cul, A., Dehaene, S., Reyes, P., Bravo, E., Slachevsky, A. (2009). Causal role of prefrontal cortex in the threshold for access to consciousness. *Brain*, **132**(9), 2531–40.

Diedrichsen, J., Shadmehr, R. (2005). Detecting and adjusting for artifacts in fMRI time series data. *Neuroimage*, **27**(3), 624–34.

Dunlosky, J., Metcalfe, J. (2009). *Metacognition*. London: Sage Publications.

Fleming, S.M., Dolan, R.J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **367**(1594), 1338–49.

Fleming, S.M., Huijgen, J., Dolan, R.J. (2012). Prefrontal contributions to metacognition in perceptual decision making. *The Journal of Neuroscience*, **32**(18), 6117–25.

Fleming, S.M., Lau, H.C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, **8**, 433.

Fleming, S.M., Ryu, J., Golfinos, J.G., Blackmon, K.E. (2014). Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain*, **137**(10), 2811–22.

Fleming, S.M., Weil, R.S., Nagy, Z., Dolan, R.J., Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science*, **329**(5998), 1541–3.

Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.P., Frith, C.D., Frackowiak, R.S.J. (1994). Statistical parametric maps in functional imaging: a general linear approach. *Human Brain Mapping*, **2**(4), 189–210.

Grill-Spector, K., Henson, R., Martin, A. (2006). Repetition and the brain: neural models of stimulus-specific effects. *Trends in Cognitive Sciences*, **10**(1), 14–23.

Green, D.M., Swets, J.A. (1966). *Signal Detection Theory and Psychophysics*. New York: Wiley.

Hanley, J.A., McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**(1), 29–36.

Hoorens, V. (1993). Self-enhancement and Superiority Biases in Social Comparison. *European Review of Social Psychology*, **4**(1), 113–39. DOI: 10.1080/14792779343000040.

Haruno, M., Kawato, M. (2006). Different neural correlates of reward expectation and reward expectation error in the putamen and caudate nucleus during stimulus-action-reward association learning. *Journal of Neurophysiology*, **95**(2), 948–59.

Hsu, M., Bhatt, M., Adolphs, R., Tranel, D., Camerer, C.F. (2005). Neural systems responding to degrees of uncertainty in human decision-making. *Science*, **310**(5754), 1680–3.

Kanske, P., Heissler, J., Schönfelder, S., Forneck, J., Wessa, M. (2013). Neural correlates of emotional distractibility in bipolar disorder, unaffected relatives and individuals with hypomanic personality. *American Journal of Psychiatry*, **170**, 1487–96.

Kanske, P., Böckler, A., Trautwein, F.-M., Singer, T. (2015). Dissecting the social brain: Introducing the EmpaToM to reveal distinct neural networks and brain-behavior relations for empathy and Theory of Mind. *Neuroimage*, **122**, 6–19.

Kim, H., Cabeza, R. (2009). Common and specific brain regions in high-versus low-confidence recognition memory. *Brain Research*, **1282**, 103–13.

Krain, A.L., Wilson, A.M., Arbuckle, R., Castellanos, F.X., Milham, M.P. (2006). Distinct neural mechanisms of risk and ambiguity: a meta-analysis of decision-making. *Neuroimage*, **32**(1), 477–84.

Kringelbach, M.L., Rolls, E.T. (2004). The functional neuroanatomy of the human orbitofrontal cortex: evidence from neuroimaging and neuropsychology. *Progress in Neurobiology*, **72**(5), 341–72.

Maniscalco, B., Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, **21**(1), 422–30.

McCurdy, L.Y., Maniscalco, B., Metcalfe, J., Liu, K.Y., de Lange, F.P., Lau, H. (2013). Anatomical coupling between distinct metacognitive systems for memory and visual perception. *The Journal of Neuroscience*, **33**(5), 1897–906.

Moore, D.A., Healy, P.J. (2008). The trouble with overconfidence. *Psychological Review*, **115**(2), 502–17.

Moritz, S., Gläscher, J., Sommer, T., Büchel, C., Braus, D.F. (2006). Neural correlates of memory confidence. *Neuroimage*, **33**(4), 1188–93.

Neubauer, A.C., Fink, A. (2009). Intelligence and neural efficiency. *Neuroscience and Biobehavioral Reviews*, **33**(7), 1004–23.

Pannu, J.K., Kaszniak, A.W., Rapcsak, S.Z. (2005). Metamemory for faces following frontal lobe damage. *Journal of the International Neuropsychological Society*, **11**, 668–76.

Potvin, P., Turmel, É, Masson, S. (2014). Linking neuroscientific research on decision making to the educational context of novice students assigned to a multiple-choice scientific task involving common misconceptions about electrical circuits. *Frontiers in Human Neuroscience*, **8**, 14.

Rounis, E., Maniscalco, B., Rothwell, J.C., Passingham, R.E., Lau, H. (2010). Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. *Cognitive Neuroscience*, **1**(3), 165–75.

Semendeferi, K., Armstrong, E., Schleicher, A., Zilles, K., Van Hoesen, G. W. (2001). Prefrontal cortex in humans and apes: a comparative study of area 10. *American Journal of Physical Anthropology*, **114**(3), 224–41.

Singer, T., Kok, B.E., Bornemann, B., Bolz, M., Bochow, C.A. (2015). *The ReSource Project. Background, Design, Samples, and Measurements. Leipzig: Max Planck Institute for Human Cognitive and Brain Sciences*. ISBN: 978-3-941 504-49-3.

Tsujimoto, S., Genovesio, A., Wise, S.P. (2010). Evaluating self-generated decisions in frontal pole cortex of monkeys. *Nature Neuroscience*, **13**(1), 120–6.

Volz, K.G., Schubotz, R.I., von Cramon, D.Y. (2005). Variants of uncertainty in decision-making and their neural correlates. *Brain Research Bulletin*, **67**(5), 403–12.

Wessa, M., Heissler, J., Schönfelder, S., Kanske, P. (2013). Goal-directed behavior under emotional distraction is preserved by enhanced task-specific activation. *Social Cognitive and Affective Neuroscience*, **8**, 305–12.

Yarkoni, T. (2009). Big correlations in little studies: inflated fMRI correlations reflect low statistical power—Commentary on Vul et al.(2009). *Perspectives on Psychological Science*, **4**(3), 294–8.

Yokoyama, O., Miura, N., Watanabe, J., *et al.* (2010). Right frontopolar cortex activity correlates with reliability of retrospective rating of confidence in short-term recognition memory performance. *Neuroscience Research*, **68**(3), 199–206.