# Confidence in Detection and Discrimination: an fMRI Study

Matan Mazor, Karl J. Friston and Stephen M. Fleming

## Objective

The current study aims to compare the brain processes that govern perceptual discrimination and detection, and the neural mechanisms that allow for metacognitive evaluations of these processes.

A fundamental property that distinguishes detection from discrimination tasks is the asymmetry in the availability of evidence for 'yes' and for 'no' responses. While discrimination requires a comparison between the relative evidence for different options, in a detection setting evidence can only be available for the presence of a stimulus and not for its absence. Conceptually, this means that confidence in the absence of a stimulus cannot rely on the magnitude of evidence for its absence and may rely instead on counterfactual reasoning regarding the likelihood of the stimulus to be detected had it been presented. Behaviorally, this difference is reflected in general lower confidence and in a weaker association between objective accuracy and subjective confidence for 'no' responses (Kanai, Walsh, & Tseng, 2010; Meuwese, van Loon, Lamme, & Fahrenfort, 2014) in detection but also in detection-like tasks (such as recognition memory; Higham, Perfect, & Bruno, 2009).

It is still unknown what are the brain mechanisms that give rise to these behavioral differences. In previous studies comparing structural and functional correlates of metacognitive sensitivity ratings across domains (Mccurdy, Maniscalco, Metcalfe, & De Lange, 2013; Morales, Lau, & Fleming, 2018), care was taken to equate task requirements and avoid the asymmetry inherent to true detection tasks. For example, instead of asking participants to perform 'old'/'new' recognition judgments, participants were asked to answer which of two presented stimuli is old. Similarly, instead of asking participants whether they detected a signal or not, a 2 Interval Forced Choice (2IFC) approach is often preferred, where participants are asked to report whether the signal was presented in the first or the second interval.

Here we wish to compare detection and discrimination within the same low-level perceptual task, while controlling for differences in task performance. The objectives of this study are:

1. Replicate previous findings of a generic confidence signal in the activity of medial prefrontal cortex (De Martino, Fleming, Garrett, & Dolan, 2013; Morales et al., 2018).
2. Test for an interaction between confidence level and task (detection/discrimination) in the BOLD response, specifically in the prefrontal cortex.
3. Within detection, test for an interaction between confidence level and response (yes/no) in the BOLD response, specifically in the prefrontal cortex and in regions that have previously been associated with counterfactual reasoning (Boorman, Behrens, & Rushworth, 2011; Neubert, Mars, Thomas, Sallet, & Rushworth, 2014).
4. Test for relationships between fluctuations in metacognitive adequacy (a trial-by-trial measure of metacognitive sensitivity; Wokke, Cleeremans, & Ridderinkhof, 2016) and the BOLD signal separately for detection and for discrimination, and for yes and no responses within detection.
5. Replicate previous findings of inter-subject correlations of functional properties of the lateral prefrontal cortex (lPFC) with metacognitive efficiency (meta-d’/d’; Fleming & Lau, 2014) in discrimination ( Yokoyama et al., 2010).
6. Identify inter-subject functional correlates of metacognitive efficiency in detection. Specifically, we will be interested to see if metacognitive efficiency in detection is predicted by different brain structures than metacognitive efficiency in discrimination.

## Design

We will test 35 healthy subjects in a 3 Tesla MRI scanner in the Welcome Centre for Human Neuroimaging, Institute of Neurology, University College London.

Participants will be acquainted with the task in a preceding behavioural session. During this session, task difficulty will be adjusted independently for detection and for discrimination, targeting around 70 % accuracy on both tasks.

The scanning session will start with a structural MP-RAGE scan, followed by a fieldmap scan. During this time, participants will perform a staircase procedure, similar to the one they performed in the preceding behavioural session, to calibrate their performance on both tasks inside the MRI scanner. After completing the calibration phase, participants will undergo 5 ten-minute functional scanner runs, each comprising one detection and one discrimination block of 40 trials each, presented in random order. After a temporally jittered rest period of 500-4000 milliseconds, the trial will begin with a fixation cross (500 milliseconds), followed by a presentation of the target for 33 milliseconds. In discrimination trials, the target will be a circle of diameter 3° containing randomly generated white noise, merged with a sinusoidal grating (2 cycles per degree; oriented 45° or -45°). In half of the detection trials, targets will not contain a sinusoidal grating and will consist of random noise only. After the offset of the stimuli, participants will use their right-hand index and middle fingers to make a perceptual decision about the orientation of the grating (discrimination blocks), or about the presence or absence of a grating (detection blocks). The response mapping will be counterbalanced between blocks, such that an index finger press will be used to indicate a clockwise tilt on half of the trials, and a counterclockwise tilt on the other half. Similarly, in half of the detection trials the index finger will be mapped to a 'yes' response, and on the other half to a 'no' response. Participants will then use their left-hand thumb to rate their confidence in their decision on a 6-point scale. The perceptual decision and the confidence rating phases will be restricted to 1500 and 2500 milliseconds, respectively. No feedback will be delivered to subjects about their performance.

Grating contrast will be fixed within each experimental block, to avoid stimulus-driven fluctuations in confidence. Nevertheless, following experimental blocks with markedly bad () or good () accuracy, grating contrast will be adjusted for the next block of the same task (divided or multiplied by a factor 0.9, for bad and good performance, accordingly). Finally, grating contrast will be adjusted for both tasks following runs in which the difference in performance between the two tasks exceeded 16.25% (contrast will be multiplied by the square root of 0.9 for the easier task and divided by the square root of 0.9 for the more difficult task).

#### Confidence rating

Subjects will rate their confidence on a 6-point scale by using two keys to increase and decrease their reported confidence level with their left-hand thumb. Confidence levels will be indicated by the size and color of a circle presented at the center of the screen. The initial size and color of the circle will be determined randomly at the beginning of the confidence rating phase, to make the number of button presses independent from the final confidence rating. The mapping between color and size to confidence will be counterbalanced between participants: for half of the participants, high confidence will be mapped to small, red circles, and for the other half, high confidence will be mapped to large, blue circles. This counterbalancing is employed to isolate confidence-related activations from activations that originate from the perceptual properties of the confidence scale or from differences in the motor requirement to press the upper and lower buttons.

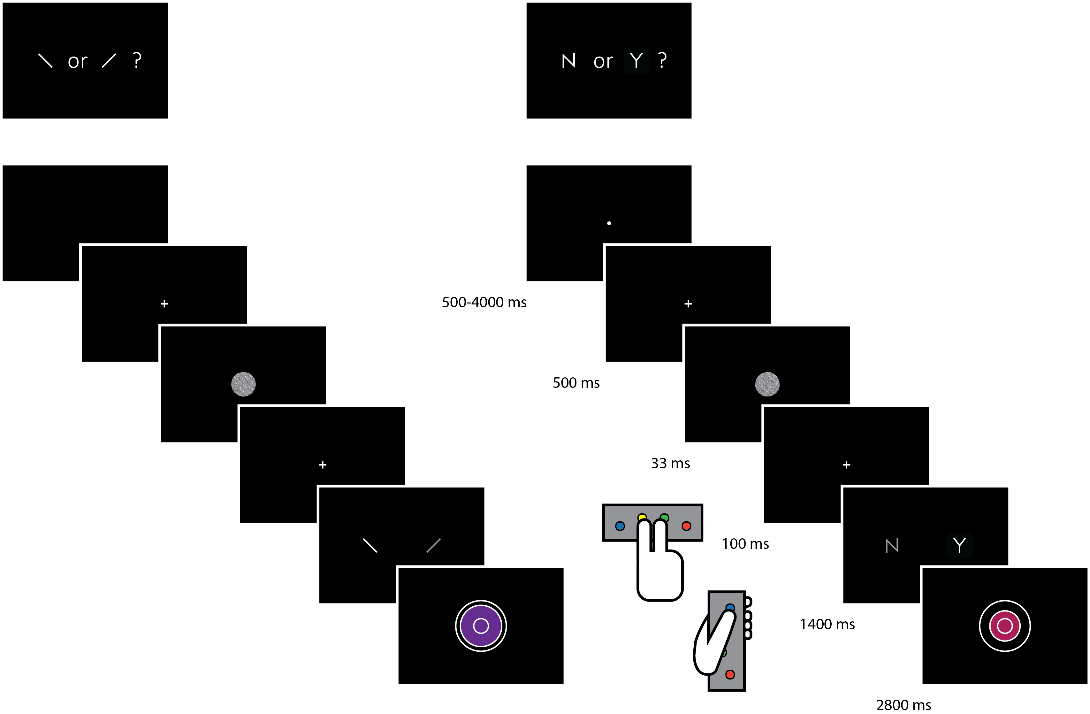


Figure 1: Experimental design for discrimination and for detection trials. Perceptual decisions are reported using the right index and middle fingers, and confidence ratings are reported using the left thumb. Confidence rating will be made by varying the size and color of a circle, with 6 options ranging from small and red to big and blue. For half of the subjects, high confidence will be mapped to the small, red circle. For the other half, high confidence will be mapped to the big, blue circle. The initial size and color of the circle will be determined randomly at the beginning of the confidence rating phase, to make the number of button presses independent from the final confidence rating.

## Scanning Parameters

Scanning will take place at the Wellcome Centre for Human Neuroimaging, London. We will use a Siemens Prisma MRI scanner with a 64-channel head coil.

We will acquire structural images using an MPRAGE sequence, followed by a double-echo FLASH (gradient echo) sequence with TE1=10ms and TE2=12.46ms )64 slices, slice thickness = 2mm, gap = 1mm, in plane FoV= 192×192 mm2, resolution = 3×3 mm2) used later for field inhomogeneity correction.

Functional scans will be acquired using a 2D EPI sequence, optimized for regions near the orbito-frontal cortex (3.0x3.0x3.0mm voxels, TR=3.36 seconds, TE = 30 ms, 48 slices tilted by -30 degrees with respect to the T>C axis, matrix size = 64x72, Z-shim=-1.4).

## Analysis

### fMRI data preprocessing

Data preprocessing will follow the procedure described in Morales and colleagues (2018):

*Imaging analysis will be performed using SPM12 (Statistical Parametric Mapping; www.fil.ion.ucl.ac.uk/spm). The first five volumes of each run will be discarded to allow for T1 stabilization. Functional images will be realigned and unwarped using local field maps (Andersson et al., 2001) and then slice-time corrected (Sladky et al., 2011). Each participant’s structural image will be segmented into gray matter, white matter, CSF, bone, soft tissue, and air/background images using a nonlinear deformation field to map it onto template tissue probability maps (Ashburner and Friston, 2005). This mapping will be applied to both structural and functional images to create normalized images to Montreal Neurological Institute (MNI) space. Normalized images will be spatially smoothed using a Gaussian kernel (6 mm FWHM). We set a within-run 1 mm rotation and 4 mm affine motion cutoff criterion.*

Preprocessing and construction of first- and second-level models will use standardized pipelines and scripts available at https://github.com/metacoglab/MetaLabCore/

### Exclusion Criteria

#### Subject exclusion

Subjects will be excluded from all analyses in the following cases:

1. They missed more than 20% of the trials.
2. Their mean accuracy in one of the tasks was lower than 60%.
3. They exceeded the head motion cutoff criterion in more than 2 experimental runs.
4. They were heavily biased toward a particular response in one of the tasks, i.e., used the same response in more than 75% of the trials.

Subjects will be excluded from any confidence-based analysis in the following cases:

1. They used the same confidence level for more than 80% of all trials.
2. For a particular response, they used the same confidence level for more than 80% of the trials.

#### Run exclusion

Individual experimental runs will not be analysed in the following cases:

1. More than 20% of the trials in the run were missed.
2. Mean accuracy in one of the tasks was lower than 60%.
3. Head motion cutoff criterion was exceeded.
4. There was a heavy bias toward one response in one of the tasks, i.e., the participant used the same response in more than 75% of the trials.

Experimental runs will not be used in the confidence analysis if:

1. The same confidence level was used for more than 95% of all trials.
2. For a particular response, the same confidence level was reported for more than 95% of the trials.

### Regions of Interest

In addition to an exploratory whole-brain analysis (corrected for multiple comparisons at the cluster level), our analysis will focus on the following *a priori* regions of interest, largely following the ROIs used at Stephen M. Fleming, van der Putten, & Daw, 2018:

1. *Frontopolar cortex* (FPC, defined anatomically). We will use a connectivity-based parcellation (Neubert et al., 2014) to define a general FPC region of interest as the total area spanned by areas FPl, FPm and BA46. The right hemisphere mask will be mirrored to create a bilateral mask.
2. *Ventromedial prefrontal cortex* (vmPFC). The vmPFC ROI will be defined as a 8-mm sphere around MNI coordinates [0,46,-7], obtained from a meta-analysis of subjective-value related activations (Bartra, McGuire, & Kable, 2013)and aligned to the cortical midline.
3. *Bilateral ventral striatum.* The ventral striatum ROIs will be specified anatomically from the Oxford-Imanova Striatal Strctural Atlas included with FSL (http://fsl.fmrib.ox.ac.uk).
4. *Posterior medial frontal cortex (pMFC).* The pMFC ROI will be defined as a 8-mm sphere around MNI coordinates [0, 17, 46], obtained from a functional MRI study on decision confidence (Stephen M Fleming, Huijgen, & Dolan, 2012)and aligned to the cortical midline.
5. *Precuneus.* The precuneus ROI will be defined as a 8-mm sphere around MNI coordinates [0,-57,18], based on Voxel Based Morphometry studies of metacognitive efficiency (Fleming et al., 2010; Mccurdy et al., 2013)and aligned to the cortical midline.

For the FPC ROI, a small-volume correction will be applied to individual voxels within the ROI for all univariate contrasts. For the multivariate analysis, we will use searchlight to scan for spatial patterns within the ROI, followed by a correction for multiple comparisons.

For all other ROIs, a GLM model will be fitted to the mean time course of voxels within the region, and multivariate analysis will be performed on all the voxels within the ROI

### Univariate Analysis

The univariate analysis will be based on a design matrix in which different trial types are modeled by different regressors (main design matrix, below). Additionally, to examine the global effect of confidence across trial types, a simpler design matrix will be fitted to the data as a first step (global confidence design matrix, below).

#### Global Confidence Design Matrix

The global confidence design matrix will consist of only 4 regressors of interest. The first two primary regressors will be 'correct trials' (trials in which the participant was correct, across tasks and responses) and 'incorrect trials' (trials in which the participant was incorrect, across tasks and responses). Single events will be modeled by a boxcar regressor with nonzero entries at the 4000 millisecond interval between the offset of the stimulus and the end of the confidence rating phase, and will be convolved with the canonical hemodynamic response function (HRF). Additionally, the design matrix will include a confidence parametric modulator for each of the first two regressors.

The construction of the regressors and the additional nuisance regressors will be handled similarly to the main design matrix.

We will use the parameter estimates from this design matrix to extract a group-level map of confidence modulation in correct responses, across all tasks and conditions.

#### Main Design Matrix

The main design matrix for the univariate GLM analysis will consist of 16 regressors of interest. There will be 8 regressors for every combination of *task x condition x response*: For example, there will be a regressor for detection trials where a signal was present ('Yes') and the subject reported seeing a signal ('Yes'; Y\_Y). The relevant trials will be modeled by a boxcar regressor with nonzero entries at the 4000 millisecond interval between the offset of the stimulus and the end of the confidence rating phase. This regressor will then be convolved with the canonical hemodynamic response function (HRF).

Each of these primary regressors will be accompanied by a parametric modulator, indicating the reported confidence for each trial. Together, this makes a total of 16 regressors:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **Task** | **Stimulus** | **Response** |
| *1* | *CW\_CW* | Discrimination | Clockwise | Clockwise |
| *2* | *CW\_CCW\_conf* |
| *3* | *CW\_CCW* | Discrimination | Clockwise | Counterclockwise |
| *4* | *CW\_CCW\_conf* |
| *5* | *CW\_CW* | Discrimination | Counterclockwise | Clockwise |
| *6* | *CW\_CW\_conf* |
| *7* | *CCW\_CCW* | Discrimination | Counterclockwise | Counterclockwise |
| *8* | *CCW\_CCW\_conf* |
| *9* | *Y\_Y* | Detection | Signal | Yes |
| *10* | *Y\_Y\_conf* |
| *11* | *Y\_N* | Detection | Signal | No |
| *12* | *Y\_N\_conf* |
| *13* | *N\_Y* | Detection | Noise | Yes |
| *14* | *N\_Y\_conf* |
| *15* | *N\_N* | Detection | Noise | No |
| *16* | *N\_N\_conf* |

Blue and red cells represent correct and incorrect responses, respectively.

In addition, trials in which the participant did not respond within the 1500 millisecond time frame will be modeled by a separate regressor. The design matrix will also include a run-wise constant term regressor, an instruction-screen regressor for the beginning of each block, four motor response regressors for each of the four response buttons, motion regressors (the first derivatives of the 6 motion parameters as extracted by SPM in the head motion correction preprocessing phase) and physiological measures regressors.

Our primary analysis will focus on correct responses only. We will apply the following contrasts to the voxel-wise beta estimates:

|  |  |
| --- | --- |
| Contrast | Interpretation |
| 1. Task   (Y\_Y + N\_N) - (CW\_CW + CCW\_CCW) | Brain regions showing a main effect of task. |
| 1. Correct rejections – Hits   N\_N - Y\_Y | Brain regions showing an effect of response within the detection task. |
| 1. Confidence (correct responses only)   Y\_Y\_conf + N\_N\_conf + CW\_CW\_conf + CCW\_CCW\_conf | Brain regions showing linear modulation of reported confidence on BOLD signal. |
| 1. Confidence x task interaction (correct responses only)   (Y\_Y\_conf + N\_N\_conf) – (CW\_CW\_conf + CCW\_CCW\_conf) | Brain regions showing differential modulation of confidence as a function of task. |
| 1. Confidence in correct rejections – Confidence in hits   Y\_Y\_conf - N\_N\_conf | Brain regions showing a differential modulation of confidence as a function of response, in the detection task. |

Contrasts 3-5 will identify regions that modulate their response as a function of reported confidence in correct responses. These regions can be further classified into two groups: brain regions that represent confidence, and brain regions that exhibit differential activation as a function of transient metacognitive adequacy (Wokke et al., 2016). While the first are expected to show a similar pattern for correct and incorrect responses, the second group is expected to show opposite signs of signal modulation as a function of response accuracy. While high confidence in a correct response is more metacognitively adequate than low confidence, the opposite is true for incorrect responses, for which a low confidence rating will reflect higher metacognitive adequacy.

To this end, we will perform four additional contrasts within brain regions that showed significant modulation as a function of confidence in contrasts 4-7:

|  |  |
| --- | --- |
| 1. Confidence (incorrect responses only)   ([0 0 0 1 0 1 0 0 0 0 0 1 0 1 0 0]) | This contrast will only be performed within regions that showed statistically significant modulation of confidence in contrast number 4. |
| 1. Confidence-task interaction (incorrect responses only)   ([0 0 0 1 0 1 0 0 0 0 0 -1 0 -1 0 0]) | This contrast will only be performed within regions that showed statistically significant modulation of confidence in contrast number 5. |
| 1. Confidence in misses – Confidence in false positives   ([0 0 0 0 0 0 0 0 0 0 0 1 0 -1 0 0]) | This contrast will only be performed within regions that showed statistically significant modulation of confidence in contrast number 6. |
| 1. Confidence in clockwise tilt – Confidence in counterclockwise tilt (incorrect responses only)   ([0 0 0 1 0 -1 0 0 0 0 0 0 0 0 0 0]) | This contrast will only be performed within regions that showed statistically significant modulation of confidence in contrast number 7. |

### Between-subject correlations

Subject-specific contrast maps of activity correlating with confidence in correct responses in discrimination ([0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0]) will be correlated against metacognitive efficiency scores for discrimination. Similarly, subject-specific contrast maps of activity correlating with confidence in correct responses in detection ([0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1]) will be correlated against metacognitive efficiency scores in detection. Here also the 5 a-priori ROIs will be analyzed in addition to a whole-brain analysis corrected for multiple comparisons.

### Multivariate analysis

Multi-voxel pattern analysis (Norman, Polyn, Detre, & Haxby, 2006) will be used to test for consistent spatial patterns in the fMRI data. We will use The Decoding Toolbox (Hebart, Gargen, & Haynes, 2015) and follow the procedures described by Morales and colleagues (2018).

|  |  |  |
| --- | --- | --- |
| Train | Test | Interpretation |
| High confidence vs. Low confidence (within correct trials) | High confidence vs. Low confidence  (within correct trials) | Spatially multivariate signal predicting confidence reports. |
| Within discrimination correct trials: high confidence vs. Low confidence | Within discrimination correct trials: high confidence vs. low confidence | Spatially multivariate signal predicting confidence reports in discrimination. |
| Within detection correct trials: high confidence vs. Low confidence | Within detection correct trials: high confidence vs. low confidence | Spatially multivariate signal predicting confidence reports in detection. |
| Within discrimination correct trials: high confidence vs. Low confidence | Within detection correct trials: high confidence vs. low confidence | Task invariant multivariate signal predicting confidence reports in detection and discrimination. |
| Within detection correct trials: high confidence vs. low confidence | Within discrimination correct trials: high confidence vs. Low confidence |
| Within detection correct trials: hits vs. misses | Within detection correct trials: hits vs. misses | Spatially multivariate signal predicting response in detection signal trials. |
| Within detection correct trials: Hits vs. Misses | Within discrimination trials: high vs. low confidence | Spatially multivariate signal predicting visibility judgment across the two tasks. |

A contrast between the last two maps will highlight brain regions that are specifically involved in inference about presence and absence, rather than judgment of available evidence.

Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage*, *76*, 412–427. https://doi.org/10.1016/j.neuroimage.2013.02.063

Boorman, E. D., Behrens, T. E., & Rushworth, M. F. (2011). Counterfactual Choice and Learning in a Neural Network Centered on Human Lateral Frontopolar Cortex. *PLoS Biology*, *9*(6), e1001093. https://doi.org/10.1371/journal.pbio.1001093

De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nature Neuroscience*, *16*(1), 105–110. https://doi.org/10.1038/nn.3279

Fleming, S. M., Huijgen, J., & Dolan, R. J. (2012). Prefrontal contributions to metacognition in perceptual decision making. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *32*(18), 6117–6125. https://doi.org/10.1523/JNEUROSCI.6489-11.2012

Fleming, S. M., van der Putten, E. J., & Daw, N. D. (2018). Neural mediators of changes of mind about perceptual decisions. *Nature Neuroscience*, *21*(4), 617–624. https://doi.org/10.1038/s41593-018-0104-6

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2009). Relating Introspective Accuracy to Individual Differences in Brain Structure. *Science*, *324*(5928), 759–764. https://doi.org/10.1126/science.1169405

Hebart, M. N., GÃ¶rgen, K., & Haynes, J.-D. (2015). The Decoding Toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. *Frontiers in Neuroinformatics*, *8*, 88. https://doi.org/10.3389/fninf.2014.00088

Higham, P. A., Perfect, T. J., & Bruno, D. (2009). Investigating strength and frequency effects in recognition memory using type-2 signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(1), 57–80. https://doi.org/10.1037/a0013865

Kanai, R., Walsh, V., & Tseng, C. (2010). Subjective discriminability of invisibility: A framework for distinguishing perceptual and attentional failures of awareness. *Consciousness and Cognition*, *19*(4), 1045–1057. https://doi.org/10.1016/J.CONCOG.2010.06.003

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*, 422–430. https://doi.org/10.1016/j.concog.2011.09.021

Mccurdy, L. Y., Maniscalco, B., Metcalfe, J., & De Lange, F. P. (2013). Anatomical Coupling between Distinct Metacognitive Systems for Memory and Visual Perception Decoded Neurofeedback Project for development of diagnostic and therapeutic system for mental disorders View project Temporal dynamics of visual imagery View proje. https://doi.org/10.1523/JNEUROSCI.1890-12

Meuwese, J. D. I., van Loon, A. M., Lamme, V. A. F., & Fahrenfort, J. J. (2014). The subjective experience of object recognition: comparing metacognition for object detection and object categorization. *Attention, Perception, & Psychophysics*, *76*(4), 1057–1068. https://doi.org/10.3758/s13414-014-0643-1

Morales, J., Lau, H., & Fleming, S. M. (2018). Domain-General and Domain-Specific Patterns of Activity Supporting Metacognition in Human Prefrontal Cortex. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *38*(14), 3534–3546. https://doi.org/10.1523/JNEUROSCI.2360-17.2018

Neubert, F.-X., Mars, R. B., Thomas, A. G., Sallet, J., & Rushworth, M. F. S. (2014). Comparison of Human Ventral Frontal Cortex Areas for Cognitive Control and Language with Areas in Monkey Frontal Cortex. *Neuron*, *81*(3), 700–713. https://doi.org/10.1016/J.NEURON.2013.11.012

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430. https://doi.org/10.1016/j.tics.2006.07.005

Wokke, M. E., Cleeremans, A., & Ridderinkhof, K. R. (2016). Sure I’m sure: Prefrontal oscillations support metacognitive monitoring of decision-making. *J. Neurosci*, *10*, 1612–1616. https://doi.org/10.1523/JNEUROSCI.1612-16.2016

Yokoyama, O., Miura, N., Watanabe, J., Takemoto, A., Uchida, S., Sugiura, M., … Nakamura, K. (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. https://doi.org/10.1016/J.NEURES.2010.07.2041