

Partially Overlapping Neural Correlates of Metacognitive Monitoring and Metacognitive Control

 Annika Boldt and Sam J Gilbert

Institute of Cognitive Neuroscience, University College London, London, WC1N 3AZ, United Kingdom

Metacognition describes the process of monitoring one's own mental states, often for the purpose of cognitive control. Previous research has investigated how metacognitive signals are generated (metacognitive monitoring), for example, when people (both female/male) judge their confidence in their decisions and memories. Research has also investigated how metacognitive signals are used to influence behavior (metacognitive control), for example, setting a reminder (i.e., cognitive offloading) for something you are not confident you will remember. However, the mapping between metacognitive monitoring and metacognitive control needs further study on a neural level. We used fMRI to investigate a delayed-intentions task with a reminder element, allowing human participants to use their metacognitive insight to engage metacognitive control. Using multivariate pattern analysis, we found that we could separately decode both monitoring and control, and, to a lesser extent, cross-classify between them. Therefore, brain patterns associated with monitoring and control are partially, but not fully, overlapping.

Key words: cognitive offloading; confidence; control; decoding; metacognition; reminders

Significance Statement

Models of metacognition commonly distinguish between monitoring (how metacognition is formed) and control (how metacognition is used for behavioral regulation). Research into these facets of metacognition has often happened in isolation. Here, we provide a study which directly investigates the mapping between metacognitive monitoring and metacognitive control at a neural level. We applied multivariate pattern analysis to fMRI data from a novel task in which participants separately rated their confidence (metacognitive monitoring) and how much they would like to use a reminder (metacognitive control). We find support for the notion that the two aspects of metacognition overlap partially but not fully. We argue that future research should focus on how different metacognitive signals are selected for control.

Introduction

Our brains possess a remarkable ability to monitor performance and to then use metacognition to control future behavior. For example, if you have low confidence that you will remember a delayed intention (metacognitive monitoring [MetaM]) like regular medication intake, you might set a reminder on your phone

(metacognitive control [MetaC]). This distinction between monitoring and control is found in the seminal metamemory framework by Nelson and Narens (1990; see also Flavell, 1976; Kluwe, 1982; Brown, 1987; Yeung et al., 2004; Efklides, 2008; Fletcher and Carruthers, 2012; Shea et al., 2014; S. M. Fleming and Daw, 2017; Fleur et al., 2021), which proposes that cognition functions at two distinct levels: the object and the meta level (Fig. 1A). Information at the object level about decisions, memories, attention, action, and so forth is re-represented at the meta level via a process of MetaM. Meanwhile, information at the meta level controls processing at the object level (MetaC). Shimamura's (2000) dynamic filtering theory extends the framework by Nelson and Narens (1990), ascribing the role of the object level to posterior cortical regions and the role of the meta level to PFC. The information flow between these regions forms the basis of MetaM and MetaC.

We are only slowly beginning to understand the neural mapping between MetaM and MetaC. This mapping or link describes the relationship that exists between MetaM and MetaC on a functional level: are these labels describing the identical process or two different computations with different

Received June 25, 2021; revised Jan. 10, 2022; accepted Jan. 12, 2022.

Author contributions: A.B. and S.J.G. designed research; A.B. and S.J.G. performed research; A.B. and S.J.G. analyzed data; A.B. wrote the first draft of the paper; A.B. and S.J.G. edited the paper; A.B. and S.J.G. wrote the paper.

This work was supported by the Wellcome Trust, who awarded Sir Henry Wellcome Postdoctoral Fellowship 206480/Z/17/Z to A.B.; and the Economic & Social Research Council, who awarded Research Grant ES/N018621/1 to S.J.G. Neither of these funding bodies played a role in the conceptualization, design, data collection, analysis, decision to publish, or preparation of the manuscript. For the purpose of Open Access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission. We thank the MetaOffloading laboratory for help with the MRI data collection; Pei-Chun Tsai for help with anatomical labeling; and Carsten Allefeld and Kai G6rgen for useful discussions regarding the data analysis.

The authors declare no competing financial interests.

Correspondence should be addressed to Annika Boldt at a.boldt@ucl.ac.uk.

<https://doi.org/10.1523/JNEUROSCI.1326-21.2022>

Copyright © 2022 the authors

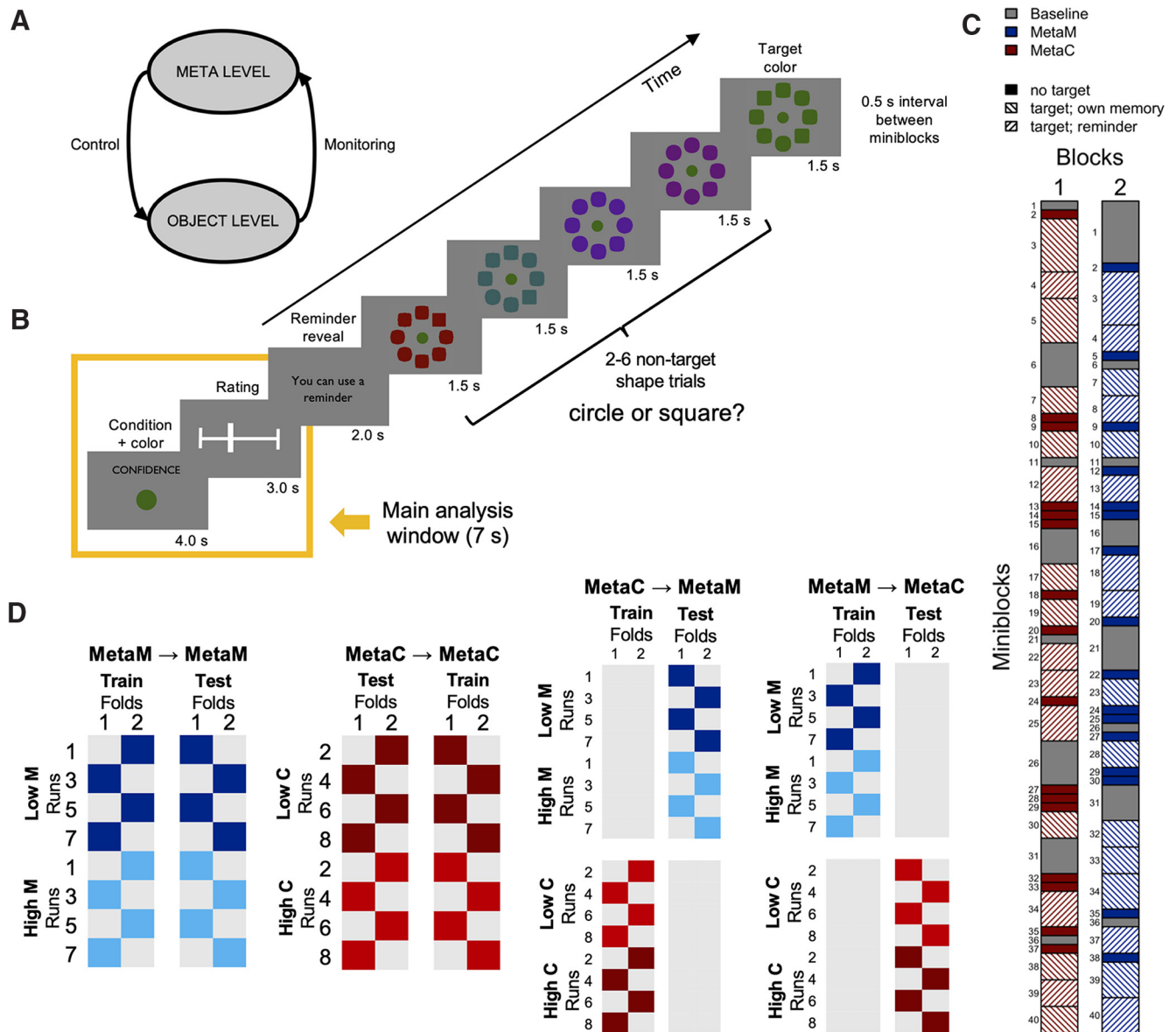


Figure 1. *A*, Model of metamemory proposed by Nelson and Narens (1990). Arrows indicate the flow of information. *B*, Example of a typical MetaM miniblock: Participants were presented with a target color and had to rate how confident they were that they would remember this color later on. It was then revealed whether or not they were allowed to use a reminder for the current miniblock (in this case, the fixation dot took on the target color for the duration of the ongoing task). The ongoing task was a shape discrimination task where participants had to judge whether an array of colored shapes was on average a circle or square. The miniblock ended unpredictably with the target color, which participants were instructed to respond to using a different key. The analysis window for the fMRI analyses is highlighted in yellow. *C*, Two example blocks showing how participants were alternatingly presented with one Baseline miniblock and then four miniblocks of the current metamemory rating condition (MetaM or MetaC), shown in different colors. The height of the rectangles reflects the approximate length of the miniblocks, their shading, and angle of the offloading condition. *D*, Design matrices for the within-category classifications (first two panels from the right) and cross-classifications (second two panels from the right). Lighter colors represent beta images modeling the lower half of ratings in the block in question. Darker colors represent higher ratings. In this example, the participant began the experiment with the MetaM condition, but approximately half of our sample started in the opposite order for balancing reasons. An inverse relationship between MetaC and MetaM is expected in the cross-classification analysis; hence, the ordering of high versus low MetaC (dark vs light red) has been flipped in the cross-classifications.

inputs? This question is important because one rationale for studying MetaM is that it can provide insight into MetaC (e.g., Boldt and Yeung, 2015; Bang and Fleming, 2018; Gherman and Philastides, 2018; Miyamoto et al., 2018; Odegaard et al., 2018; Shekhar and Rahnev, 2018; Ye et al., 2018; Masset et al., 2020; Wokke et al., 2020). This would be strengthened if the mapping between the two were better understood. Furthermore, dissociations have been found between MetaM and MetaC. For example, in some circumstances, young children (Redshaw et al., 2018), obsessive-compulsive disorder patients (Vaghi et al., 2017), older adults (Dunlosky and Connor, 1997), and individuals with autism spectrum conditions (Grainger et al., 2016)

have a diminished mapping between MetaM and MetaC, which could lead to suboptimal behavioral regulation. However, the potential neural substrates for this variability are unknown.

One of the reasons why the MetaM-MetaC mapping has received little attention is that the two aspects of metamemory are usually studied in isolation (but see Koriati et al., 2006, 2014; Son and Schwartz, 2009; Qiu et al., 2018; Mei et al., 2020; Schulz et al., 2021). Studies on MetaM commonly explore the variables that affect how confident people feel and the associated neural correlates. For example, neuroimaging studies have identified a widespread network of involved regions, including the rostrolateral PFC (Yokoyama et al., 2010; S. M. Fleming et al., 2012;

Allen et al., 2017) and also the precuneus specifically for meta-memory studies (e.g., Baird et al., 2013; McCurdy et al., 2013; Ye et al., 2018). Moreover, machine-learning techniques have been used to “decode” brain patterns associated with low versus high confidence, using both fMRI (Cortese et al., 2016; Hebart et al., 2016; Morales et al., 2018) and EEG (Boldt and Yeung, 2015). Research on MetaC, on the other hand, has focused on situations in which metacognitive experiences are used for learning, communication, or speed-accuracy trade-off, to name a few (e.g., Metcalfe and Finn, 2008; Bahrami et al., 2010; Shea et al., 2014; Guggenmos et al., 2016; Desender et al., 2019; Lak et al., 2020; Frömer et al., 2021).

Most of what we know about the link between monitoring and control comes from the field of cognitive control and error monitoring. Electrophysiological correlates have been found that signal not only when an error has been committed but are also sensitive to correct-trial performance fluctuations (Allain et al., 2004; Yeung et al., 2004). Such monitoring of errors often results in lower response speed immediately after a mistake, a robust and often-replicated phenomenon termed post-error slowing (Rabbitt, 1966; Notebaert et al., 2009; Danielmeier and Ullsperger, 2011). In addition to errors, conflict signals appear to be monitored by the posterior medial frontal cortex, including the dorsal anterior cingulate cortex (dACC). The lateral PFC (laPFC) is thought to receive this input and implement cognitive control (Ridderinkhof et al., 2004). It should be noted that participants are often not aware of such errors or response conflicts and that these studies are not directly measuring metacognitive signals. Nevertheless, evidence from this domain suggests that similar brain regions support MetaM and MetaC. Qiu et al. (2018) conducted four elegant fMRI experiments, using a decision–redecision paradigm: Participants were presented twice in a row with each stimulus and rated both their response and confidence for each presentation. They reasoned that participants would engage MetaM for their initial response and use MetaC to revise and improve decisions in the redecision phase. Their analyses revealed an involvement of dACC in the first response and IFPC in the second. However, because the order of the decision–redecision phases was always the same, it is impossible to conclude whether the redecision phase really triggered more MetaC or whether the signal observed in IFPC was instead a “late” monitoring one. Another open question is whether MetaM and MetaC rely on similar representations.

In order to address these questions, it is necessary to study both aspects of metacognition in a single paradigm, which we did using a cognitive offloading task. Cognitive offloading is the use of physical action to reduce cognitive demand (e.g., setting external reminders rather than relying on internal memory). Previous research has demonstrated a MetaM–MetaC link whereby individuals are more likely to set reminders (MetaC) when they have low confidence in their memory abilities over and above the influence of their actual memory performance (MetaM) (Dunn and Risko, 2016; Risko and Gilbert, 2016; Hu et al., 2019; for a review see Gilbert et al., 2022). This finding is a robust pattern that can even be observed when reminder setting is not explicitly instructed (Boldt and Gilbert, 2019) or when confidence was measured in an unrelated perceptual task (Gilbert, 2015). Here, we use a decoding approach to examine this link at a neural level.

Participants performed a delayed intention task where in separate blocks they engaged in MetaM (how confident am I that I will remember?) or MetaC (how much would I like a reminder?). This allowed us to answer two questions: (1) Do similar brain patterns characterize MetaM and MetaC? If so, (2) can the neural

patterns that characterize specific acts of MetaC be exhaustively characterized in terms of their associated processes of MetaM? We answered these questions by examining cross-classification between MetaM and MetaC: the extent to which a classifier trained on one judgment can decode the other. Insofar as this is possible, this implies a shared neural code for MetaM and MetaC. But if cross-classification is weaker than decoding MetaM and MetaC individually, this implies that their neural bases do not overlap fully.

Materials and Methods

Participants. We trained 29 participants in a behavioral task during a first session. After reviewing their training data, 22 participants returned to the laboratory for a second MRI session 1–21 d later, excluding 7 participants (2 unsuited for MRI because of safety regulations, 2 had extreme staircase values, 3 were unavailable for a second session). Another participant was excluded after scanning because of excessive movement in the scanner. This resulted in a final sample of 21 participants, of which 15 were female and 6 were male. While we determined our sample size based on practical constraints and on available resources, the final sample size of $N = 21$ is nevertheless in accordance with previous MRI studies using similar methods (Hebart et al., 2016; Morales et al., 2018; Qiu et al., 2018). Participants were 20.3 years on average (18–26 years) and were paid £36 for their participation in both sessions (~90 and 150 min). All participants were right-handed, had intact color vision, no uncorrected visual impairments, and had not been diagnosed with any psychiatric or neurologic disorders. All testing was approved by the local ethics committee, and participants gave informed consent before taking part in the study.

Experimental design. In order to investigate the extent to which neural patterns associated with MetaM and MetaC are similar or distinct, we had to study both aspects of metacognition within a single paradigm. Participants underwent short miniblocks of ongoing shape discrimination trials. For this ongoing task, participants had to quickly and accurately decide whether an array of colored shapes grouped around a fixation dot looked on average more like a circle or a square (De Gardelle and Summerfield, 2011) by pressing one of two buttons. The response categories were equally likely. During some of these miniblocks, participants also had to maintain a delayed intention to press a different button if the stimulus appeared in a target color (Fig. 1B). Participants were allowed to use reminders (cognitive offloading) to support their prospective memory in approximately half of the miniblocks, which meant that the central fixation dot of the stimulus took on the target color for the duration of the miniblock. Instead of having to rely on their memory, participants could then simply wait for the color of the shapes to match the color of the fixation dot, making the fulfillment of the delayed intention much easier. There were 12 colors, placed equidistant in RGB space. Within each miniblock, colors were drawn without replacement. There was only one target color per miniblock, presented at the beginning of the miniblock, and its occurrence during the ongoing-task trials always terminated the miniblock.

The task comprised three within-subject experimental conditions (20% Baseline, 40% MetaM, and 40% MetaC; Table 1) each structured into miniblocks. A miniblock comprised presentation of a target color (except for in the Baseline condition, which had no prospective memory element), a single metacognitive rating or cursor placement, followed by 3–7 ongoing-task trials. The number of trials per miniblock was drawn from an exponential distribution with a mean of $\mu = 1.1$; in other words, shorter miniblocks were more frequent than longer miniblocks. Each of the eight blocks consisted of 94 shape trials spread unevenly across 40 miniblocks (Fig. 1C). The critical difference between our two key conditions was the metacognitive rating given about the target color before each miniblock. In the MetaC condition, participants reported how much they would want to set a reminder to help them remember this target color. The higher the rating given by the participant, the greater the likelihood of receiving a reminder, which occurred on ~50% of miniblocks. More specifically, ratings larger than the moving median of

Table 1. List of experimental conditions

	Baseline	MetaM	MetaC
Proportion	20% (32 partial + 32 full = 64 miniblocks)	40% (64 partial + 64 full = 128 miniblocks)	40% (64 partial + 64 full = 128 miniblocks)
Delayed intention	—	Target color	Target color
Rating	Cursor placement	Very unconfident to very confident	Sure reminder to sure own memory
Reminders	—	50% (random)	Based on moving median rating cut-off

the past 20 MetaC ratings were assigned a reminder, whereas ratings below this cutoff had to be solved using only unaided memory. In the MetaM condition, participants reported their prospective confidence in remembering the target color. However, this had no influence on the likelihood of receiving a reminder, which occurred on a randomly selected 50% of miniblocks. In other words, the two conditions also differed in the relationship between participants' ratings and the provision of reminders. In the MetaM condition, participants' ratings had no influence on whether or not they received a reminder. In the MetaC condition, on the other hand, which miniblock contained a reminder was largely determined by participants' ratings. Therefore, in the MetaM condition, participants engaged in MetaM but did not exercise MetaC. In the MetaC condition, they exercised control to make a decision, which is known to be guided by metacognition (Boldt and Gilbert, 2019; Gilbert, 2015). However, they were not explicitly asked to make a direct metacognitive judgment.

In the Baseline condition, there was no target color and thus no prospective-memory component (and no need for a reminder). The rating participants were asked to give was thus an "empty" one, that is a scale without labels but with a cursor was presented on screen together with two little markers indicating where the cursor should be placed on the scale. Participants then had to move the cursor to the indicated position. In all three conditions, participants were instructed to move the cursor at least once to submit a rating.

Each block was comprised of only two of the three conditions, the Baseline condition together with either the MetaM or the MetaC condition and alternated between the two. Within each block, conditions were predictable, that is they always followed the order of one Baseline miniblock followed by four other miniblocks. We determined the optimal order of conditions using simulations, allowing us to maximize the efficiency of our design. The main analysis window was the initial 7 s of the task (presentation of target color and rating). At the time of these prospective ratings, participants were still unaware whether or not they would receive a reminder, keeping our key contrast free of confounds, which would have been unavoidable had we chosen a retrospective confidence judgment as is more commonly used in the field. To increase the number of instances this analysis window was shown, we therefore included partial miniblocks; that is, half of the time (20 miniblocks per run), the miniblock ended immediately after the rating without the need to perform any shape classification trials or search for the target.

The study comprised two sessions. The purpose of the first session was assessment of MRI safety, completion of a prestudy questionnaire on how much participants liked the 12 colors used in the task, and training in the behavioral task (presented in MATLAB using Psychtoolbox3) (Kleiner et al., 2007). Participants first completed eight practice blocks, each introducing them to a new aspect of the paradigm. They then completed four experimental runs that were identical to the task they would have to complete while in the scanner, each lasting ~9 min. During the second session, participants first underwent two practice blocks outside of the scanner (each lasting ~5 min) to remind them of the task before they completed eight runs in the scanner, with a 6 min T1 scan between the fourth and fifth run. One participant only completed six blocks because of feeling unwell inside the scanner. Because of the unbalanced design, we decided to exclude this participant from all multivariate analyses.

At the end of the second session, participants were furthermore asked to fill in a post-experiment questionnaire, asking them to rate the liking of all colors again, together with how difficult they found them and several additional questions to determine whether they perceived the MetaM and MetaC conditions as similar, how much control they felt

during these conditions, how they used the reminders depending on whether or not they asked for them, and how they approached each rating. The orientation of the rating scales was flipped halfway through the experiment to avoid confounding visuomotor processes with low versus high ratings. The order of scale orientations, response keys for the shape task, and the order of the conditions were counterbalanced across participants.

MRI data collection and preprocessing. We used a 1.5T Siemens Avanto scanner with a 32-channel head coil and MRI-safe button boxes. We acquired both T1-weighted structural images, as well as T2*-weighted EPI (64 × 64; 3.2 × 3.2 × 3.2 mm voxels) with BOLD contrast. We used a multiband acquisition sequence with acceleration factor = 3, TE = 54.8 ms, flip angle = 75°, to record 39 interleaved, axial slices (3.2 mm thick, oriented approximately to the anterior commissure–posterior commissure plane). This allowed us to cover most of the brain with an effective repetition time of 1.3 s per volume. Encoding phase direction was anterior to posterior. Functional scans were acquired in eight runs, each comprising 410 volumes (~9 min). The first five volumes in each session were discarded to allow for T1 equilibration effects. Between the fourth and fifth functional scans, an ~6 min T1-weighted MPRAGE structural scan was collected.

All preprocessing was done using SPM12 (<https://www.fil.ion.ucl.ac.uk/spm/software/spm12/>). The T1-weighted images were skull stripped, and their origin was set to the anterior commissure. We then realigned the EPI volumes and normalized them into 3 mm cubic voxels with fourth-degree B-spline interpolation using normalization parameters derived from segmentation of the coregistered structural scan, then smoothed with an isotropic 8 mm FWHM Gaussian kernel.

Statistical analysis. Analyses of behavioral data were conducted using R version 3.6.0 (Planting of a Tree) with the additional packages *plyr*, *plotrix*, *Hmisc*, *R.matlab*, *viridis*, *effsize*, *raincloudplots*, *ggplot2*, *grid*, *gridExtra*, and *Rmisc*. Statistical tests were conducted two-sided if not stated otherwise. For *t* tests, we reported effect sizes as Cohen's *d*, and for ANOVAs as partial eta square, η^2_p . For the fMRI analyses, the volumes acquired during the eight sessions were treated as separate time series. For each time series, the variance in the BOLD signal was decomposed with a set of regressors in a GLM. Three regressors were generated to code for the target color presentation and the rating as a 7s boxcar, separately for miniblock and rating conditions (Baseline in MetaM blocks, low MetaM rating, high MetaM rating in MetaM blocks and Baseline in MetaC blocks, low MetaC rating, high MetaC rating in MetaC blocks). Six additional regressors were generated that represented effects of no interest, specifically, stimulus presentation as a stick function, separately for targets and nontargets, the ongoing task spanning from the onset of the first to the last shape stimulus of the miniblock, separately for whether there was a prospective-memory requirement (Baseline vs MetaM and MetaC) and the time when the computer revealed to the participant whether they were allowed to use a reminder as a stick function, separately for Reminder and Own Memory miniblocks. All regressors were convolved with a canonical hemodynamic response function. The regressors outlined above, along with six regressors representing residual movement-related artifacts and the mean over scans, comprised the full model for each session. The data and model were high-pass filtered at a cutoff of 1/128 Hz. Parameter estimates for each regressor were calculated from the least mean squares fit of the model to the data. Effects of interest were assessed in a random-effect analysis by first forming subject-specific contrasts subtracting the Baseline from the other two conditions. The resulting contrast images were entered into a repeated-measures ANOVA using nonsphericity correction (Friston et al., 2002), representing a condition agnostic

selection contrast to identify a network of regions active in the rating task. Results are reported applying a height threshold of $p < 0.001$ uncorrected in conjunction with an extent threshold determined by SPM12 to achieve $p < 0.05$ familywise error correction for multiple comparisons across the whole-brain volume. ROI analyses were conducted by extracting subject-specific contrast estimates from the resulting ROIs with the toolbox MarsBaR (Brett et al., 2002), then entering the resulting data into an ANOVA in R using the same correction procedure described above.

The logic behind the key analysis of our study was the following: Replicating and extending previous findings (Cortese et al., 2016; Hebart et al., 2016; Morales et al., 2018), we first trained separate classifiers to detect (1) whether participants were in a high or low confidence state (MetaM), and (2) whether they had high or low desire for a reminder (MetaC). These classifiers could then also be combined in a cross-classification analysis, that is, whether a classifier trained on MetaM ratings can also predict MetaC ratings (and vice versa). Insofar as this cross-classification is possible, this suggests shared brain representations for both aspects of metacognition. Going one step further, we then compared within-category classification to cross-classification accuracy to distinguish between two possible patterns of results: If MetaM and MetaC are based on the exact same representational code, there should be no difference in classification accuracy. If, on the other hand, MetaM and MetaC share partially overlapping patterns, we should find significantly higher classification accuracy for within- than cross-category classification, but significantly-different-from-zero accuracy for cross-classification.

For the multivariate-pattern analyses, we used The Decoding Toolbox (Hebart et al., 2015), based on the beta images resulting from the previously described general linear models (except that the models were refit to unsmoothed, unnormalized data and the MetaM and MetaC boxcar regressors were split into two regressors each using a median split on the respective metacognitive rating). When we ran our four separate decoding analyses, two drew the training and testing data from the same condition (low vs high ratings for the MetaM and MetaC conditions, respectively; defined by block-, condition-, and subject-wise median splits), whereas the other two cross-classified (train on low vs high MetaM ratings and test on high vs low MetaC ratings and vice versa; the rating scale had to be flipped for MetaC as low confidence implies high desire for a reminder). For each of these analyses, a linear support vector machine was trained to discriminate between low versus high ratings given the patterns of BOLD activity across voxels. Given the alternating block design and the fact that the orientation of the scale was flipped halfway through the study, we had two low and two high rating images available for each training or testing fold, resulting in a twofold procedure (Fig. 1D).

We used a whole-brain searchlight approach (Kriegeskorte et al., 2006), meaning that for each voxel a separate support vector machine was built, fitted to the beta values within a sphere with a radius of 3 voxels (9.6 mm). This resulted in three-dimensional decoding accuracy maps in native space for each participant and analyses. Decoding accuracy is calculated relative to chance level (subtracted by 50%, so a 5% accuracy corresponds to 55%). These maps were then normalized into MNI space (using the same normalization parameters as the univariate analyses) and smoothed using a Gaussian kernel (FWHM, 4 mm). This kernel was half of the one used for the univariate analyses. This was done to avoid excessive smoothing, given that the searchlight analysis already imposes spatial smoothing on the data. The resulting images were entered into a one-sample t test using SPM12. This allowed assessment of voxels showing consistently higher decoding accuracy in a random-effect analysis. We note that the suitability of second-level t tests has been challenged for information-like measures, such as classification accuracy, where classifier performance can meaningfully be above, but not below, chance levels (Allefeld et al., 2016; Hirose, 2020). However, this characteristic does not apply to our two key hypothesis-testing analyses. For the cross-classification between MetaM and MetaC, high MetaM could either predict higher or lower MetaC. For the comparison between within- and cross-classification, accuracy for one classification could be higher or lower than the other. Therefore, in both cases, our statistical

tests are valid because they are performed on data that could meaningfully take values both above and below zero.

Along with the main multivariate pattern analyses described above, we conducted an additional analysis. Here, we used a similar approach to the univariate ROI analysis described above by defining a condition agnostic contrast (the mean of all four decoding analyses), extracting ROIs with significantly above-chance decoding accuracy and then entering the resulting classification accuracies into a repeated-measures ANOVA with factors ROI, training condition (MetaM/MetaC), and classification type (within-condition/cross-classification).

Results

Behavioral results

Our sample included 22 participants, one of which was excluded because of excessive motion in the MRI scanner (for more details, see Materials and Methods). Participants performed the tasks with a high level of accuracy (mean shape-discrimination accuracy = 93.4%, SEM = 0.84%; nonsignificant shape bias, $t_{(20)} = 1.2$, $p = 0.25$, $d = 0.26$; mean target-detection rate = 88.1%, SEM = 3.37%; NB chance target-detection accuracy would be 8.3%; false alarm rate = 0.8%). With our design, we decided against using a direct manipulation of difficulty (e.g., spacing some colors closer to each other in color space) as this would have made it difficult to interpret any effect of confidence because of its inherent confound with a difficulty manipulation. Instead, we relied on natural fluctuations in confidence caused, for example, by individual preferences for colors or fatigue. Figure 2A shows that average, unaided memory performance varied across colors with some colors (e.g., the fourth color, a shade of green) being associated with lower accuracy when participants had to remember this target color unaided by a reminder. Moreover, this figure shows that not all participants had the same inherent color-difficulty profile and that instead some participants perceived particular colors as more difficult than others. Performance in the Baseline condition was high. Here, an indicator of compliance with instructions is participants' placement of the cursor between two thin lines marked on the scale (Fig. 2B). Participants reported that those lines were difficult to see in the scanning session. Nevertheless, their cursor locations peaked around the marked location and landed within the marked positions on approximately half the trials ($\delta = 47.6\%$).

We next established that the reminders aided participants in their fulfillment of the delayed intentions by comparing target-detection error rates for miniblocks in which participants had to use their own memory (fixation dot stayed white) to miniblocks in which they were allowed to use a reminder (fixation dot took on target color), shown in Figure 2C. In both conditions, error rates were reduced when reminders could be used ($F_{(1,20)} = 20.5$, $p < 0.001$, $\eta^2_p = 0.51$; $t > 3.4$, $p < 0.01$, $d > 0.37$ when tested separately for the MetaM and MetaC conditions). Error rates did not differ significantly between conditions ($F < 1$), nor was there an interaction between the two factors ($F < 1$).

When asked explicitly after the experiment how similar they perceived the two conditions, participants rated the conditions as similar but not identical (mean = 0.68 on a scale from 0 = totally different to 1 = exactly the same; min = 0.28; max = 0.98). Indeed, we found that participants' perception of the two conditions differed in how much control participants felt they had over the reminders. On a scale ranging from 0 = no control to 1 = full control, participants rated the MetaM condition with a mean of 0.32 (min = 0.00; max = 0.88) and the MetaC condition with a mean of 0.80 (min = 0.06; max = 0.98). This difference was significant ($t_{(20)} = 6.4$; $p < 0.001$, $d = 1.94$). This shows that participants were

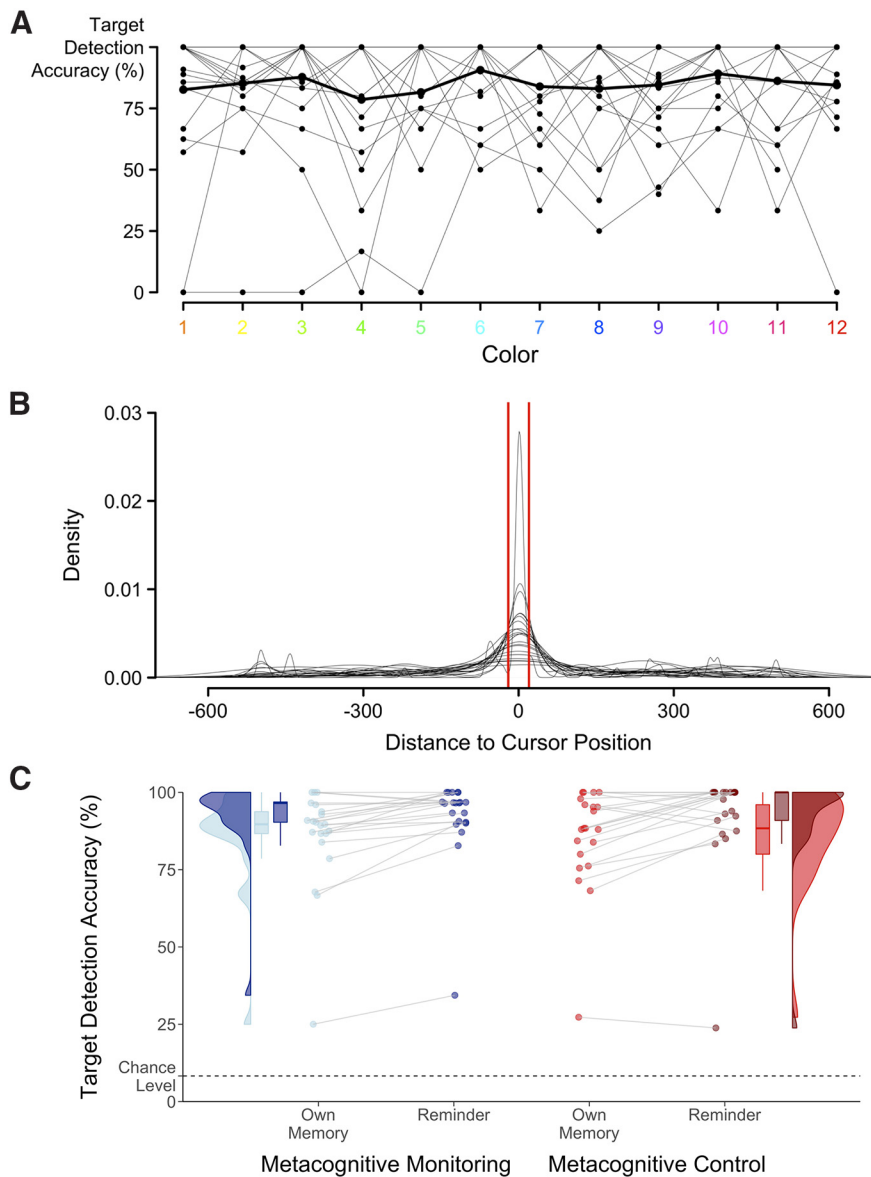


Figure 2. *A*, Target detection accuracy across the range of 12 target colors, placed equidistant in RGB space. Only trials without a reminder were included, and the data were averaged across conditions (MetaM and MetaC). Thick black line indicates the sample average. Thinner lines indicate individual participants. *B*, Placement of the cursor in the Baseline condition relative to the marked position on the scale (shown in red). Each black line indicates the data from a single participant. *C*, Target detection accuracy as a function of miniblock condition (MetaM and MetaC) and offloading condition (Own Memory and Reminder) as a raincloud plot (Allen et al., 2019). The distributions depict the densities of the participant-wise accuracy averages for each condition. Boxplots represent the median and interquartile range of the data. Lines indicate the individual observations.

able to grasp the key difference that distinguished the two conditions.

We furthermore aimed to rule out that any condition differences found in the pattern classification analyses could be caused by behavioral differences in how the different ratings were approached. First, Figure 3*A* shows that the average ratings participants gave for each individual color were almost indistinguishable whether they were giving a metacognitive-monitoring or metacognitive-control rating. Indeed, if we correlated the average ratings for each color for each individual participant, there was an average relationship of $r = 0.76$ with 19 of 21 participants showing a significant, positive relationship between the MetaM and the MetaC rating for different colors. Relatedly, participants' rating and rating RT distributions for the two types of ratings were closely matched

(Fig. 3*B,C*). It is important to note that participants did not receive any instructions to use these scales in the same way (except for being asked to use the entire range of the scale in both cases).

Furthermore, neither of the metacognitive rating conditions showed a systematic relationship between confidence and accuracy: For retrospective confidence judgments, it is commonly found that these correlate; that is, participants express lower confidence on errors than on correct trials (confidence resolution or Type II sensitivity). In the MetaC condition, on the other hand, participants' ratings triggered reminders, so we would expect to see the opposite pattern: Trials for which they expressed a high need for a reminder should naturally be the ones on which they were allowed to offload and error rates should therefore be lower. However, we found no significant difference between correct- and error-trial ratings in any of the four conditions (MetaC reminder, $t_{(14)} = 0.2$, $p = 0.88$, $d = 0.04$; MetaC own memory, $t_{(19)} = 0.1$, $p = 0.95$, $d = 0.02$; MetaM reminder, $t_{(17)} = 1.0$, $p = 0.34$, $d = 0.30$; MetaM own memory, $t_{(20)} = 1.1$, $p = 0.28$, $d = 0.23$; participants with missing data excluded from the respective analysis). We furthermore correlated the dichotomous accuracy vector with our continuous confidence measure for all four data cells, separately for each participant. The distributions of these correlations are shown in the right panels of Figure 3*D*. None was significantly different from zero ($t < 1.0$, $p > 0.32$). Together, both the prospective nature of the ratings in the present task (i.e., participants might have felt they needed to invest more into trials in which they felt less confident or wanted a reminder more) and our unique offloading design could potentially have led to a reduced confidence resolution, but this was the case for both rating conditions.

Univariate fMRI results

We first performed univariate analyses to identify brain regions activated by the requirement to encode new intentions and make metacognitive judgements about them. We therefore averaged across the two metacognition conditions (MetaM and MetaC) and compared them with the Baseline condition, allowing us to find ROIs activated by our task. After familywise error correcting for multiple comparisons, this contrast revealed seven regions showing increased BOLD signal in the metacognitive conditions (Table 2; Fig. 4*A,B*).

Within the seven ROIs, activity was then compared between the metacognition conditions. More specifically, activity was extracted in two separate contrasts (MetaM > Baseline and MetaC > Baseline) and then compared. This comparison is orthogonal to the initial selection contrast and therefore unbiased (Kriegeskorte et al., 2009). BOLD signal was higher for the MetaC

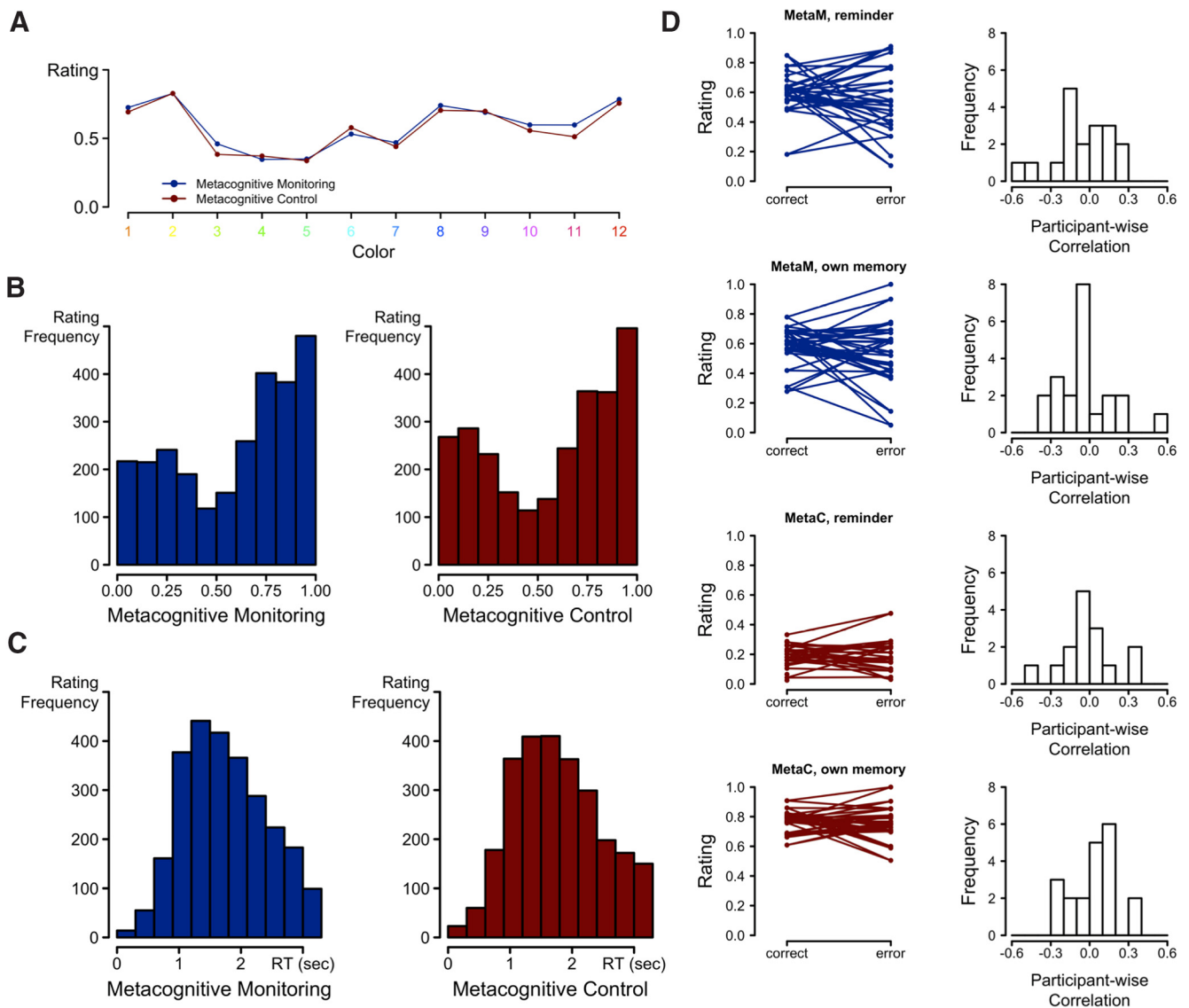


Figure 3. **A**, Participants' ratings of MetaM (low to high confidence) and MetaC (high to low need for a reminder; to match orientation of the confidence scale) showed a similar pattern across the 12 different colors placed equidistant in RGB space. **B**, **C**, Histogram of the (**B**) ratings and (**C**) rating RTs pooled across all participants as a function of rating type. **D**, Metacognitive ratings as a function of objective accuracy and condition shown as participant-wise averages (left panels) and correlations (right panels).

Table 2. Regions of increased signal in the MetaM and MetaC conditions, relative to the Baseline condition

Contrast	Label	Laterality	Peak voxel MNI coordinates	k_E	p_{FWE} cluster-corrected	Z_{max} at peak level
(MetaM + MetaC) > Baseline	Occipital and parietal cortex (calcarine cortex; cuneus; precuneus; lateral ventricles; all regions both left and right)	right and left	30, -55, 5	1338	<0.001	5.14
	Middle frontal gyrus	right	42, 32, 44	345	<0.001	4.94
	Inferior and middle frontal gyri	left	-42, 20, 26	802	<0.001	4.68
	Superior and middle frontal gyri	right	27, 62, 5	152	0.002	4.49
	Supplementary motor area	left	-6, 23, 44	117	0.009	4.27
	Angular gyrus	left	-57, -55, 44	150	0.003	4.16
	Precentral and postcentral gyri	right	18, -28, 65	75	0.046	4.08
	Angular gyrus	right	57, -58, 44	87	0.028	4.05

than the MetaM condition in all seven ROIs (Fig. 4C), and this main effect was significant when examined in an ROI (7) $\times$ Condition (2: MetaC/MetaM) repeated-measures ANOVA ($F_{(1,20)} = 8.1$, $p = 0.01$, $\eta_p^2 = 0.29$). There was furthermore a reliable main effect of ROI ($F_{(7,140)} = 7.8$, $p < 0.001$, $\eta_p^2 = 0.28$) as well as a significant interaction of the two factors ($F_{(7,140)} = 3.4$, $p < 0.01$, $\eta_p^2 = 0.14$), reflecting that the absolute signal change and also the difference in signal

change were larger in some ROIs compared with others. Together, these results show that regions that respond to the conditions requiring delayed intentions and metacognitive judgments showed higher activity when participants rated how much they would like a reminder (MetaC) compared with how confident they were (MetaM).

We repeated the univariate analyses for deactivations, revealing six “task-negative” regions showing decreased signal in the

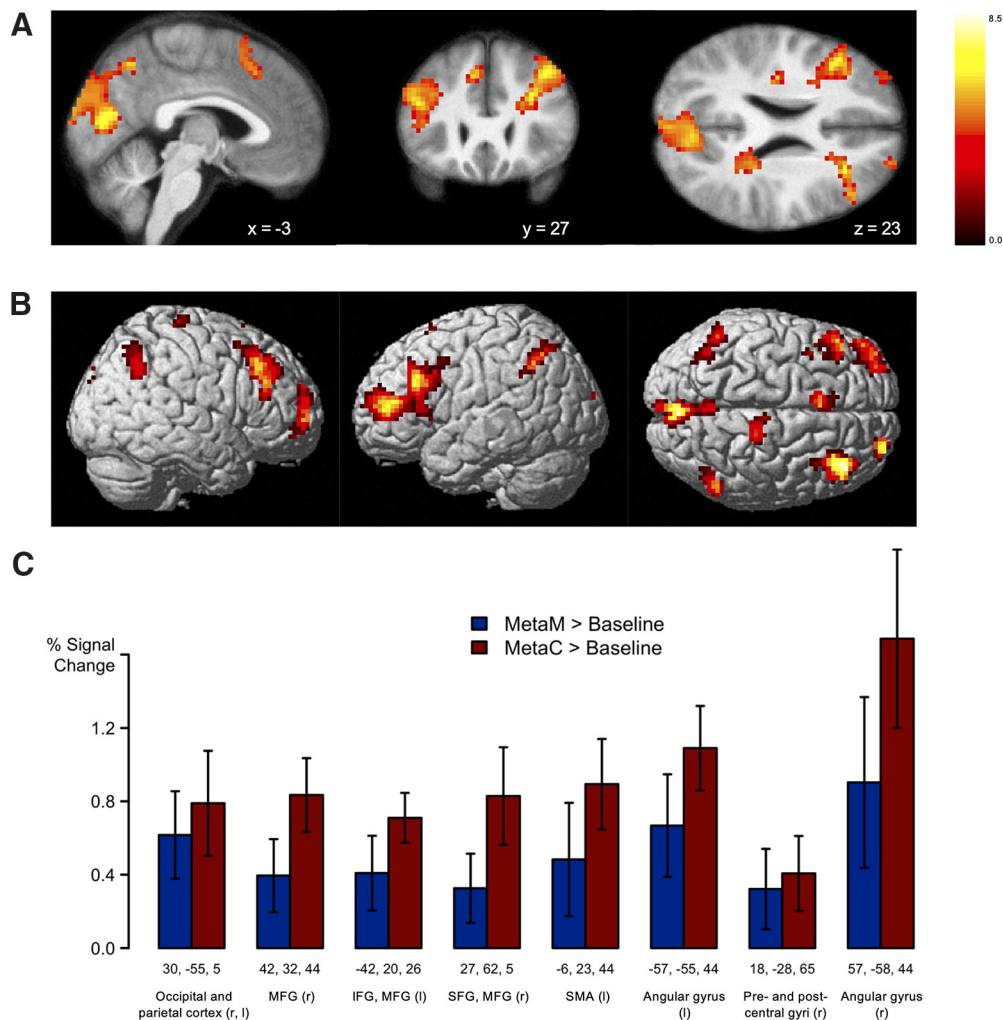


Figure 4. *A, B*, The regions of significant signal change in the contrast of (MetaM + MetaC) > Baseline. *A*, Significant results were plotted on sagittal ($x = -3$), coronal ($y = 27$), and axial ($z = 23$) views of the skull-stripped, mean, normalized structural image. *B*, Three-dimensional renderings of results on right hemisphere, left hemisphere, and superior views. *C*, Percent signal change for both metacognitive rating conditions compared with the Baseline condition, in ROIs defined by the contrasts shown in Table 2. r, Right; l, left; MFG, middle frontal gyrus; IFG, inferior frontal gyrus; SFG, superior frontal gyrus; SMA, supplementary motor area. Error bars indicate $\pm$ within-subject CIs (95%) according to Morey (2008).

conditions requiring delayed intentions and metacognitive judgments compared with Baseline (Fig. 5; Table 3). These regions included the cingulate and paracingulate cortices, supplementary motor area, supramarginal gyrus, middle and inferior temporal gyri, occipital gyri, and anterior cingulate gyrus. Within these task-negative ROIs, there was more deactivation when participants rated how confident they were (MetaM) compared with how much they would like a reminder (MetaC); however, BOLD signal did not differ significantly between the MetaC and the MetaM condition ($F_{(1,20)} = 1.3$, $p = 0.26$, $\eta_p^2 = 0.06$). There was a reliable main effect of ROI ($F_{(5,100)} = 18.2$, $p < 0.001$, $\eta_p^2 = 0.48$). The interaction was not significant ($F < 1$).

Multivariate fMRI results

The multivariate analyses allowed us to address our two key questions: (1) Do the brain patterns of different metacognitive experiences also distinguish different acts of control? (2) Can the neural patterns that characterize specific acts of MetaC be exhaustively characterized in terms of their associated metacognitive experiences? In a first analysis, we attempted to decode confidence (MetaM). Figure 6*A, B* shows the resulting decoding accuracy maps corrected for chance level and multiple comparisons, resulting in nine clusters that contained

meaningful information when predicting whether the brain was currently in a low or high confidence state, including the anterior cingulate gyrus, parietal occipital sulcus, central sulcus, superior parietal lobule, superior occipital gyrus, cuneus, precuneus, supplementary motor area, occipital fusiform gyrus, calcarine cortex, superior corona radiata, and precentral gyrus (Table 4).

We then repeated the equivalent analysis for the MetaC condition, again successfully decoding whether participants gave a low or high rating (i.e., desire for a reminder) from five clusters, including the occipital pole, lateral occipital cortex, superior parietal lobule, superior frontal gyrus (medial segment), and middle temporal gyrus (Table 4). Together, these analyses show that the neuroimaging data contain meaningful patterns that distinguish both different metacognitive experiences (low vs high confidence) and different acts of MetaC (low vs high desire for a reminder).

Having established the existence of meaningful patterns across the brain that distinguish different levels both of MetaM and MetaC, we could then ask whether it was possible to cross-classify the two aspects of metacognition. More specifically, we trained classifiers to distinguish low from high confidence beta images (MetaM) and tested them to predict high versus low

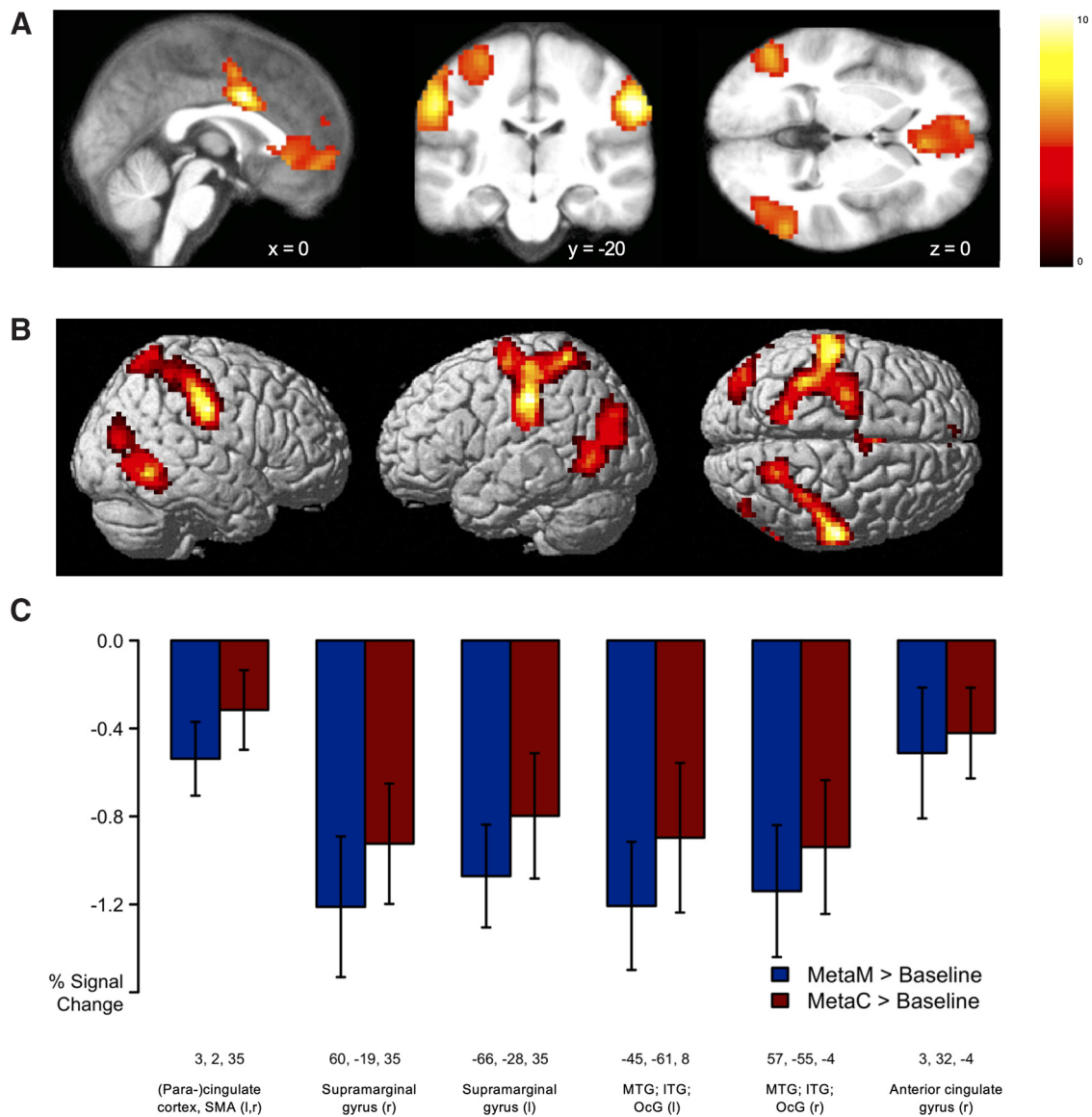


Figure 5. *A, B*, The regions of significant signal change in the contrast of Baseline > (MetaM + MetaC). *A*, Significant results were plotted on sagittal ($x = 0$), coronal ($y = -20$), and axial ($z = 0$) views of the skull-stripped, mean, normalized structural image. *B*, Three-dimensional renderings of results on right hemisphere, left hemisphere, and superior views. *C*, Percent signal change for both metacognitive rating conditions compared with the Baseline condition, in ROIs defined by the task-negative contrasts shown in Table 3. r, Right; l, left; SMA, supplementary motor area; MTG, middle temporal gyrus; ITG, inferior temporal gyrus; OcG, occipital gyri. Error bars indicate $\pm$ within-subject CIs (95%) according to Morey (2008).

Table 3. Regions of decreased signal in the MetaM and MetaC conditions, relative to the Baseline condition

Contrast	Label	Laterality	Peak voxel MNI coordinates	k_E	p_{FWE} cluster-corrected	Z_{max} at peak level
Baseline > (MetaM + MetaC)	Cingulate and paracingulate cortices; supplementary motor area; all regions both right and left	right and left	3, 2, 35	311	<0.001	5.99
	Supramarginal gyrus	right	60, -19, 35	809	<0.001	5.96
	Supramarginal gyrus	left	-66, -28, 35	1249	<0.001	5.86
	Middle and inferior temporal gyri; occipital gyri	left	-45, -61, 8	534	<0.001	4.93
	Middle and inferior temporal gyri; occipital gyri	right	57, -55, -4	508	<0.001	4.92
	Anterior cingulate gyrus	right	3, 32, -4	560	<0.001	4.61

MetaC ratings. An inverse relationship is expected between MetaM and MetaC ratings (i.e., low confidence predicts high desire for reminder, and vice versa). Therefore, one of the scales was inverted to perform this analysis. Above-chance classification accuracy can be interpreted as overlapping patterns encoding both MetaM and MetaC. The same analysis was then applied to the opposite direction (train on

MetaC, test on MetaM). Importantly, we found overlapping patterns that encode these different types of metacognitive ratings. However, only for the latter analysis direction (train on MetaC, test on MetaM) did we find above-chance classification accuracy after correcting for multiple comparisons. The surviving cluster was located in the left superior and middle frontal gyri. These findings show that brain patterns associated with different metacognitive

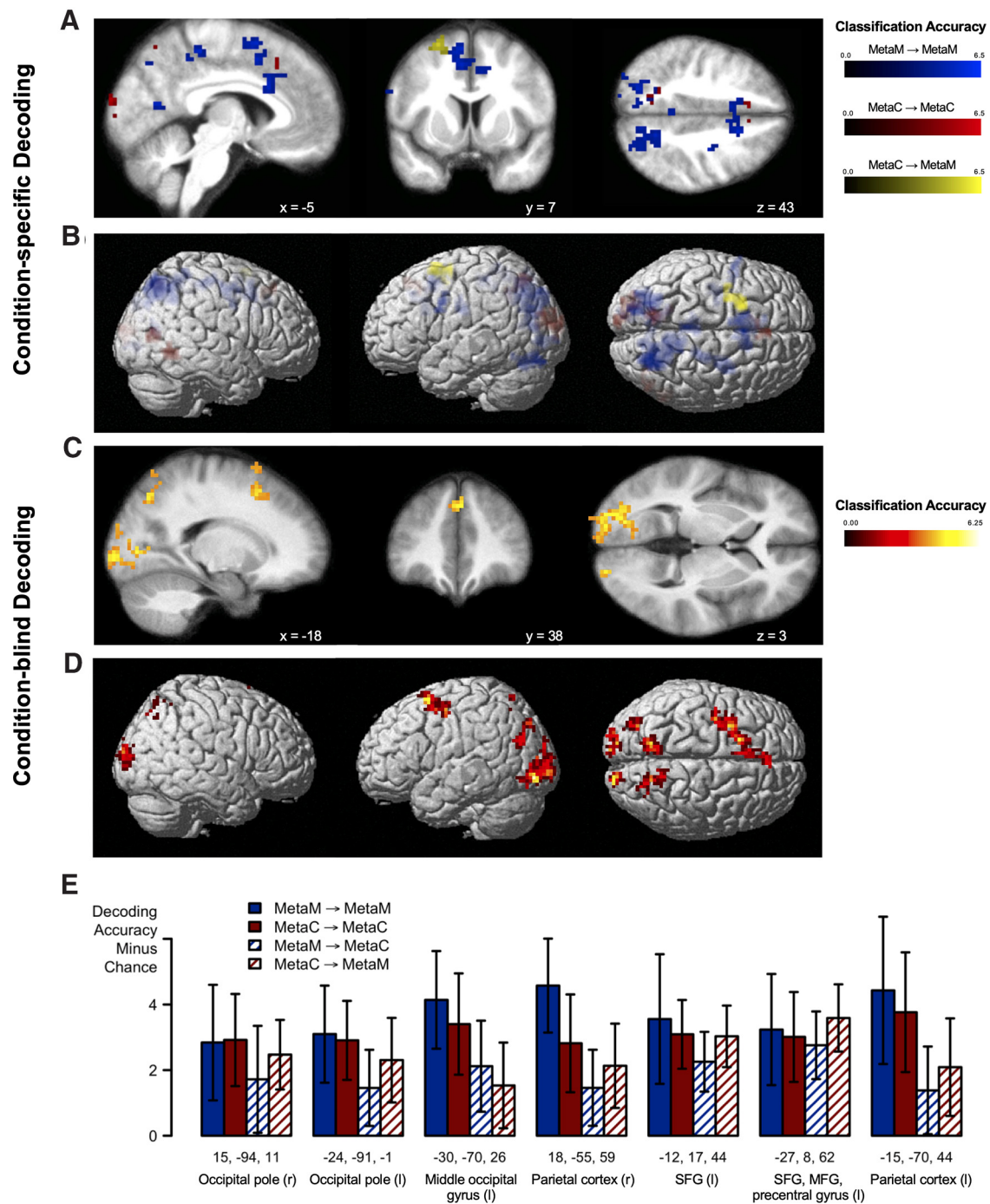


Figure 6. *A, B*, The above-chance decoding accuracy maps for the condition-specific classification analyses (blue: train on MetaM, test on MetaM; red: train on MetaC, test on MetaC; yellow: train on MetaC, test on MetaM). *A*, Significant results were plotted on sagittal ($x = -5$), coronal ($y = 7$), and axial ($z = 43$) views of the skull-stripped, mean, normalized structural image. *B*, Three-dimensional renderings of results on right hemisphere, left hemisphere, and superior views. *C, D*, The above-chance decoding accuracy maps when all four classification analyses were averaged (condition-blind decoding). *C*, Significant results were plotted on sagittal ($x = -18$), coronal ($y = 38$), and axial ($z = 3$) views of the skull-stripped, mean, normalized structural image. *D*, Three-dimensional renderings of results on right hemisphere, left hemisphere, and superior views. *E*, Above-chance classification accuracy for all four classification analyses (trained and/or tested on MetaM and MetaC, respectively) in ROIs defined by a condition-blind selection contrast that averaged across all four analyses, listed in Table 5. r, Right; l, left; SFG, superior frontal gyrus; MFG, medial frontal gyrus. Error bars indicate $\pm$ within-subject CIs (95%) according to Morey (2008).

experiences (low vs high confidence) also distinguish different acts of MetaC (low vs high desire for a reminder).

To address our second key question, we compared classification accuracy resulting from the two different types of classification analyses described above: within-category (test on MetaM and train on MetaM; test on MetaC and train on MetaC) versus across-category classification (i.e., cross-classification: test on

MetaM and train on MetaC; test on MetaC and train on MetaM). We first performed a condition-blind analysis by averaging across all four decoding analyses. This identified ROIs that contain information in one or more of the analyses in an unbiased manner, yielding significant effects in the occipital pole, middle occipital gyrus, parietal cortex (superior parietal lobule, precuneus), superior frontal gyrus, middle frontal gyrus, and

Table 4. Clusters of above-chance classification accuracy in the four classification analyses

Multivariate pattern analysis	Label	Laterality	Peak voxel MNI coordinates	k_E	p_{FWE} cluster-corrected	Z_{max} at peak level
MetaM (low vs high confidence)	Anterior cingulate gyrus	left	−3, 17, 26	58	<0.001	4.57
	Parietal occipital sulcus	left	−18, −85, 41	310	<0.001	4.51
	Central sulcus	right	21, −28, 53	90	<0.001	4.51
	Superior parietal lobule; superior occipital gyrus; cuneus; precuneus	right	24, −70, 50	404	<0.001	4.43
	Supplementary motor area (both left and right)	right	15, 14, 44	149	<0.001	4.10
	Occipital fusiform gyrus	left	−33, −67, −19	80	<0.001	3.94
	Calcarine cortex; cuneus (all regions both left and right)	left	−6, −67, 14	43	0.003	3.89
	Superior corona radiata	left	−24, −13, 32	38	0.006	3.75
	Precentral gyrus	left	−45, 2, 35	26	0.046	3.69
	Occipital pole	left	−9, −100, 14	93	<0.001	4.09
MetaC (low vs high need for a reminder)	Lateral occipital cortex	right	36, −73, 5	28	0.029	4.08
	Superior parietal lobule	left	−12, −67, 53	50	0.001	3.79
	Superior frontal gyrus (medial segment)		0, 29, 50	31	0.017	3.63
	Middle temporal gyrus	right	57, −52, −4	32	0.014	3.58
	Superior and middle frontal gyri	left	−21, 11, 62	52	<0.001	4.14
MetaC → MetaM	—	—	—	—	—	—
MetaM → MetaC	—	—	—	—	—	—

Table 5. Clusters of above-chance classification accuracy in the condition-blind classification analyses

Label	Laterality	Peak voxel MNI coordinates	k_E	p_{FWE} cluster-corrected	Z_{max} at peak level
Occipital pole	right	15, −94, 11	90	0.002	4.54
Occipital pole	left	−24, −91, −1	321	<0.001	4.51
Middle occipital gyrus	left	−30, −70, 26	89	0.002	4.47
Parietal cortex (superior parietal lobule; precuneus)	right	18, −55, 59	115	<0.001	4.22
Superior frontal gyrus	left	−12, 17, 44	114	<0.001	4.12
Superior and middle frontal gyri; precentral gyrus	left	−27, 8, 62	121	<0.001	3.96
Parietal cortex (superior parietal lobule; precuneus)	left	−15, −70, 44	91	0.002	3.65

precentral gyrus (Table 5; Fig. 6C,D). Within the resulting ROIs, classification accuracies in the four analyses could then be compared (Fig. 6E) to address the question whether decoding accuracy differed significantly between the within-condition classification and the cross-classification analyses. Taking an analogous approach to our univariate analysis, these comparisons were unbiased because they were orthogonal to the analysis used to define the ROIs. We entered the classification accuracies from these regions into a repeated-measures ANOVA with factors ROI, training condition (MetaM/MetaC), and classification type (within-condition/between-condition cross-classification). There was a significant main effect of classification type ($F_{(1,19)} = 6.2$, $p = 0.02$, $\eta_p^2 = 0.25$), with higher classification accuracy for within-condition classifications than between-condition cross-classifications. This finding can be interpreted as partially overlapping neural representations between MetaM and MetaC as opposed to perfect overlap between the patterns associated with the two aspects of metacognition. Moreover, there was no effect of the conditions on which the classifier was trained or which ROI was analyzed, F values < 1. We found a significant interaction between ROI and category (within vs between classification; $F_{(6,114)} = 2.4$, $p = 0.03$, $\eta_p^2 = 0.11$), reflecting that the difference between within-condition and across-condition decoding analyses was larger in some ROIs compared with others. No other interactions were significant (F values < 1). In sum, while our results demonstrate overlapping patterns between MetaM and MetaC, they also suggest that patterns of MetaC cannot exhaustively be characterized by associated patterns of MetaM when participants report their confidence.

Discussion

MetaM is only valuable insofar as it can subsequently influence control. And MetaC can only occur if there are metacognitive

representations to begin with, which can then be used to adjust future behavior. The two processes must therefore be intimately related, yet the mapping between them requires further study, especially on a neural level. Here we report three main findings: (1) we can separately decode MetaM and MetaC; (2) brain patterns of different levels of metacognition monitoring (low vs high confidence) also distinguish different acts of MetaC (low vs high desire for a reminder); and (3) this overlap in patterns while significant is only partial. These findings suggest that patterns of brain activity corresponding to specific acts of MetaC are partially, but not fully, characterized by associated acts of MetaM.

Our cross-classification analysis revealed involvement of the left superior and middle frontal gyri, which form part of the laPFC in both MetaM and MetaC. The role of the laPFC in metacognition has already been highlighted by previous studies, suggesting a role in domain-general metacognition (Morales et al., 2018; see also Vaccaro and Fleming, 2018), in the readout of sensory information as an input for confidence signals (Shekhar and Rahnev, 2018), and more broadly in a mediating role of more rostral parts of laPFC in metacognitive accuracy (S. M. Fleming et al., 2010; Rounis et al., 2010). Crucially, the laPFC has also been implicated in MetaC (Qiu et al., 2018; for review, see Shimamura, 2000; S. M. Fleming and Dolan, 2014; Seow et al., 2021) matching its more general proposed involvement in cognitive control (MacDonald et al., 2000; Ridderinkhof et al., 2004). Our study therefore extends this growing body of research that implies an involvement of the laPFC in metacognition and cognitive control.

Seeing as MetaC could not be characterized exhaustively in terms of the MetaM judgments we investigated, this raises the obvious question of which other signals might contribute to MetaC. We consider two main possibilities. The first possibility is that nonmetacognitive signals also play a role in influencing

MetaC. A wide variety of signals may be relevant here, such as motivation, the costs and rewards associated with different levels of performance, serial dependencies, fatigue, states of interoceptive and bodily awareness reflecting endogenous signals like arousal (Allen et al., 2016; Hauser et al., 2017; Rouault et al., 2018), and so on. This influence of nonmetacognitive signals on MetaC was already acknowledged in the seminal paper by Nelson and Narens (1990) introducing their metamemory framework. The influence of a wide variety of signals on control is also central to an influential model from the cognitive control literature, the Expected Value of Control model (EVC; Shenhav et al., 2013). This model emphasizes the flexibility with which different control signals are selected, based on the costs and benefits associated with these signals. The model proposes that the dACC integrates both costs and benefits to form the expected value of control. Seeing as MetaC may involve the integration of multiple relevant signals, including the products of MetaM and additional nonmetacognitive signals as well, this could potentially explain the greater univariate signal we observed for the MetaC than the MetaM condition. This suggests the incorporation of additional processes into the MetaC judgment beyond those involved in MetaM. We also note that the factor of within-versus cross-classification interacted significantly with region, although there was no main effect of region. This suggests that the overlap between MetaM and MetaC is greater in some regions than others.

A second possible contribution to the MetaC condition is the integration of additional metacognitive signals, beyond the confidence judgment required by the MetaM condition. In our task, for instance, participants' desire for reminder might have been influenced not only by confidence in their prospective memory but also confidence in their perceptual judgements. Consistent with this, behavioral evidence suggests that confidence judgements are influenced by a variety of domain-general and domain-specific signals (Gilbert, 2015; Kantner et al., 2018; Rouault et al., 2018). Confidence can be regarded as an explicit representation of uncertainty, and uncertainty exists at multiple levels throughout the brain (as noted by the Bayesian brain hypothesis; Knill and Pouget, 2004). Therefore, the metacognitive signals measured in the MetaM condition probably form only a subset of the metacognitive signals which may have contributed to MetaC judgments.

Our paradigm involved measurement of only a single MetaC judgment, which may have been influenced by multiple MetaM signals. In reality, there are multiple types of both MetaM and MetaC. Take, for example, the situation of a foreign language student studying for a test at her desk during the early evening hours. The student reads a word on a flashcard, and we can assume she has access to two relevant metacognitive signals: On the one hand, there is the certainty with which the word is perceived in the waning light; the other is the certainty with which the word is recognized from memory. The former confidence should guide her decision whether or not to switch on her desk lamp. The latter confidence should guide her decision whether or not to place the flashcard on the pile marked as "restudy." Similarly, the same confidence signal could lead to opposite consequences depending on the situation, as shown by Carlebach and Yeung (2020). The authors report that low confidence leads to advice-seeking when the quality of the advice is known and high. However, when the quality of the advice is unknown, people tend to seek advice, especially when they have high confidence to test the accuracy of the advisor. How does the brain then "harvest" these various confidence signals and route them

to the appropriate act(s) of MetaC? How does it flexibly switch to a different set of signals when required to do so? How are metacognitive signals weighted by past rewards, and how do such weightings shift when our goals change? Questions such as these could potentially be addressed by adapting the present paradigm to a situation involving two or more forms of MetaM and MetaC.

The key finding of our study was the cross-classification between MetaM and MetaC. At a whole-brain corrected threshold, this analysis produced a significant effect in only one direction (train on MetaC and test on MetaM). It is not clear whether this reflects an asymmetry in cross-classification, or simply a thresholding artifact. This could be an interesting question to investigate in future work. Our finding of successful cross-classification is in line with the notion that metacognition should be regarded as a cornerstone of cognitive control. Twenty years ago, this point was made prominently by Fernandez-Duque et al. (2000), who drew parallels between metacognitive and executive control functions. Similarly, Yeung and Summerfield (2012, 2014) have suggested that error monitoring, as it is commonly studied in the cognitive-control literature, constitutes an inverse, binary measure of graded confidence. It is therefore not surprising that decision confidence is tracked by a well-established electrophysiological marker of error monitoring, the error positivity (Boldt and Yeung, 2015). Other empirical examples of links between metacognition and cognitive control are the findings that metacognitive efficiency correlates with cognitive control ability (Drescher et al., 2018) and that confidence modulates the speed accuracy trade-off on a trial-by-trial basis with participants prioritizing accuracy over response speed after a previous low-confidence decision (Desender et al., 2019). The latter effect is reminiscent of post-error slowing (Rabbitt, 1966; Jentzsch and Dudschig, 2009; Danielmeier and Ullsperger, 2011), one of the most extensively studied effects of the cognitive control literature.

Our findings bear some interesting parallels to another recent decoding study. Mei et al. (2020) reported the results from two behavioral experiments, each focused on a different type of prospective decision (belief of successfully classifying a visual stimulus vs deciding whether or not to attend to the stimulus during the upcoming trial). The authors found that it was possible to use the data from one experiment (awareness ratings, confidence ratings, and accuracy in previous trials) to predict the prospective decision from the respective other experiment and vice versa. This cross-classification analysis therefore highlights similarities of MetaM (in this case: beliefs of successfully classifying the upcoming stimulus) and MetaC (in this case: decision to attend), showing that both aspects of metacognition appear in the context of the same behavioral precursors.

Despite the theoretical distinction between two binary facets of metacognition and the two different labels assigned to the conditions, the conceptual distinction between the two is not as straightforward as it may seem. For example, our MetaM condition might still be considered to involve an act of MetaC in the sense that participants need to use their metacognitive knowledge to control the act of placing the cursor on the scale to indicate low versus high confidence. We suggest that the key distinction between the conditions is that MetaM involves relatively direct read-out of metacognitive (e.g., confidence) signals, whereas MetaC involves the use of the signals to inform more complex behaviors rather than report the metacognitive experience itself. However, seeing as metacognitive reports are, at least to some degree, inferential in nature (Koriat, 1993), MetaM and

MetaC might be seen as extreme points on a continuum rather than dichotomous processes.

In conclusion, our study delineates the similarities and divisions between neural correlates of MetaM and MetaC. Ultimately, understanding the link between monitoring and control could inform interventions, such as metacognitive training in conditions including brain injury (J. Fleming et al., 2017), schizophrenia (Moritz and Woodward, 2007), and obsessive-compulsive disorder (Fisher and Wells, 2008). We propose that a full understanding of the relationship between monitoring and control will require a focus on the ways in which distinct metacognitive signals are integrated and selectively routed to appropriate acts of MetaC.

References

- Allain S, Carbonnell L, Falkenstein M, Burle B, Vidal F (2004) The modulation of the Ne-like wave on correct responses foreshadows errors. *Neurosci Lett* 372:161–166.
- Allefeld C, Görden K, Haynes JD (2016) Valid population inference for information-based imaging: from the second-level t-test to prevalence inference. *Neuroimage* 141:378–392.
- Allen M, Frank D, Schwarzkopf DS, Fardo F, Winston JS, Hauser TU, Rees G (2016) Unexpected arousal modulates the influence of sensory noise on confidence. *Elife* 5:e18103.
- Allen M, Glen JC, Müllensiefen D, Schwarzkopf DS, Fardo F, Frank D, Callaghan MF, Rees G (2017) Metacognitive ability correlates with hippocampal and prefrontal microstructure. *Neuroimage* 149:415–423.
- Allen M, Poggiali D, Whitaker K, Marshall TR, Kievit RA (2019) Raincloud plots: a multi-platform tool for robust data visualization. *Wellcome Open Res* 4:63.
- Bahrami B, Olsen K, Latham PE, Roepstorff A, Rees G, Frith CD (2010) Optimally interacting minds. *Science* 329:1081–1085.
- Baird B, Smallwood J, Gorgolewski KJ, Margulies DS (2013) Medial and lateral networks in anterior prefrontal cortex support metacognitive ability for memory and perception. *J Neurosci* 33:16657–16665.
- Bang D, Fleming SM (2018) Distinct encoding of decision confidence in human medial prefrontal cortex. *Proc Natl Acad Sci USA* 115:6082–6087.
- Boldt A, Gilbert SJ (2019) Confidence guides spontaneous cognitive offloading. *Cogn Res Princ Implic* 4:45.
- Boldt A, Yeung N (2015) Shared neural markers of decision confidence and error detection. *J Neurosci* 35:3478–3484.
- Brett M, Anton JL, Valabregue R, Poline JB (2002) Region of interest analysis using an SPM toolbox. 8th International Conference on Functional Mapping of the Human Brain. Sendai: Japan.
- Brown AL (1987) Metacognition, executive control, self-regulation, and other more mysterious mechanisms. In: *Metacognition, motivation, and understanding* (Weinert FE, Kluwe RH, eds), pp 95–116. Hillsdale, NJ: Erlbaum.
- Carlbach N, Yeung N (2020) Flexible use of confidence to guide advice requests. *PsyArXiv*.
- Cortese A, Amano K, Koizumi A, Kawato M, Lau H (2016) Multivoxel neurofeedback selectively modulates confidence without changing perceptual performance. *Nat Commun* 7:13669.
- Danielmeier C, Ullsperger M (2011) Post-error adjustments. *Front Psychol* 2:233.
- De Gardelle V, Summerfield C (2011) Robust averaging during perceptual judgment. *Proc Natl Acad Sci USA* 108:13341–13346.
- Desender K, Boldt A, Verguts T, Donner TH (2019) Confidence predicts speed-accuracy tradeoff for subsequent decisions. *Elife* 8:e43499.
- Drescher LH, Van den Bussche E, Desender K (2018) Absence without leave or leave without absence: examining the interrelations among mind wandering, metacognition and cognitive control. *PLoS One* 13:e0191639.
- Dunlosky J, Connor LT (1997) Age differences in the allocation of study time account for age differences in memory performance. *Mem Cognit* 25:691–700.
- Dunn TL, Risko EF (2016) Toward a metacognitive account of cognitive offloading. *Cogn Sci* 40:1080–1127.
- Efklides A (2008) Metacognition. *Eur Psychologist* 13:277–287.
- Fernandez-Duque D, Baird JA, Posner MI (2000) Executive attention and metacognitive regulation. *Conscious Cogn* 9:288–307.
- Fisher PL, Wells A (2008) Metacognitive therapy for obsessive-compulsive disorder: a case series. *J Behav Ther Exp Psychiatry* 39:117–132.
- Flavell JH (1976) Metacognitive aspects of problem solving. In: *The nature of intelligence* (Resnick LB, ed), pp 231–236. Hillsdale, NJ: Erlbaum.
- Fleming J, Ownsworth T, Doig E, Hutton L, Griffin J, Kendall M, Shum DH (2017) The efficacy of prospective memory rehabilitation plus metacognitive skills training for adults with traumatic brain injury: study protocol for a randomized controlled trial. *Trials* 18:3–11.
- Fleming SM, Daw ND (2017) Self-evaluation of decision-making: a general Bayesian framework for metacognitive computation. *Psychol Rev* 124:91–114.
- Fleming SM, Dolan RJ (2014) The neural basis of metacognitive ability. In: *The cognitive neuroscience of metacognition*, pp 245–265. Berlin: Springer.
- Fleming SM, Weil RS, Nagy Z, Dolan RJ, Rees G (2010) Relating introspective accuracy to individual differences in brain structure. *Science* 329:1541–1543.
- Fleming SM, Huijgen J, Dolan RJ (2012) Prefrontal contributions to metacognition in perceptual decision making. *J Neurosci* 32:6117–6125.
- Fletcher L, Carruthers P (2012) Metacognition and reasoning. *Philos Trans R Soc Lond B Biol Sci* 367:1366–1378.
- Fleur DS, Bredeweg B, van den Bos W (2021) Metacognition: ideas and insights from neuro- and educational sciences. *NPJ Sci Learn* 6:13.
- Friston KJ, Glaser DE, Henson RN, Kiebel S, Phillips C, Ashburner J (2002) Classical and Bayesian inference in neuroimaging: applications. *Neuroimage* 16:484–512.
- Frömer R, Nassar MR, Bruckner R, Stürmer B, Sommer W, Yeung N (2021) Response-based outcome predictions and confidence regulate feedback processing and learning. *Elife* 10:e62825.
- Gherman S, Philastides M (2018) Human VMPFC encodes early signatures of confidence in perceptual decisions. *Elife* 7:e38293.
- Gilbert S, Boldt A, Sachdeva C, Scarampi C, Tsai P-C (2022) Outsourcing memory to external tools: A review of ‘intention offloading’. <https://doi.org/10.31234/osf.io/ahqtz>.
- Gilbert SJ (2015) Strategic use of reminders: influence of both domain-general and task-specific metacognitive confidence, independent of objective memory ability. *Conscious Cogn* 33:245–260.
- Grainger C, Williams DM, Lind SE (2016) Metacognitive monitoring and control processes in children with autism spectrum disorder: diminished judgement of confidence accuracy. *Conscious Cogn* 42:65–74.
- Guggenmos M, Wilbertz G, Hebart MN, Sterzer P (2016) Mesolimbic confidence signals guide perceptual learning in the absence of external feedback. *Elife* 5:e13388.
- Hauser TU, Allen M, Purg N, Moutoussis M, Rees G, Dolan RJ (2017) Noradrenaline blockade specifically enhances metacognitive performance. *Elife* 6:e24901.
- Hebart MN, Görden K, Haynes JD (2015) The Decoding Toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. *Front Neuroinform* 8:88.
- Hebart MN, Schriever Y, Donner TH, Haynes JD (2016) The relationship between perceptual decision variables and confidence in the human brain. *Cereb Cortex* 26:118–130.
- Hirose S (2020) Valid and powerful group statistics for decoding accuracy: information prevalence inference using the i-th order statistic i-test. *BioRxiv* 578930.
- Hu X, Luo L, Fleming SM (2019) A role for metamemory in cognitive offloading. *Cognition* 193:104012.
- Jentsch I, Dudschig C (2009) Why do we slow down after an error? Mechanisms underlying the effects of posterror slowing. *Q J Exp Psychol (Hove)* 62:209–218.
- Kantner J, Solinger LA, Grybinas D, Dobbins IG (2018) Confidence carry-over during interleaved memory and perception judgments. *Mem Cognit* 47:195–211.
- Kleiner M, Brainard D, Pelli D, Ingling A, Murray R, Broussard C (2007) What's new in Psychtoolbox-3. *Perception* 36:1–16.
- Kluwe RH (1982) Cognitive knowledge and executive control. In: *Human mind–animal mind* (Griffin D, ed), pp 201–224. New York: Springer.
- Knill DC, Pouget A (2004) The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci* 27:712–719.

- Koriat A (1993) How do we know that we know? The accessibility model of the feeling of knowing. *Psychol Rev* 100:609–639.
- Koriat A, Ma'ayan H, Nussinson R (2006) The intricate relationships between monitoring and control in metacognition: lessons for the cause-and-effect relation between subjective experience and behavior. *J Exp Psychol Gen* 135:36–69.
- Koriat A, Ackerman R, Adiv S, Lockl K, Schneider W (2014) The effects of goal-driven and data-driven regulation on metacognitive monitoring during learning: a developmental perspective. *J Exp Psychol Gen* 143:386–403.
- Kriegeskorte N, Goebel R, Bandettini P (2006) Information-based functional brain mapping. *Proc Natl Acad Sci USA* 103:3863–3868.
- Kriegeskorte N, Simmons WK, Bellgowan PS, Baker CI (2009) Circular analysis in systems neuroscience: the dangers of double dipping. *Nat Neurosci* 12:535–540.
- Lak A, Okun M, Moss MM, Gurnani H, Farrell K, Wells MJ, Reddy CB, Kepecs A, Harris KD, Carandini M (2020) Dopaminergic and prefrontal basis of learning from sensory confidence and reward value. *Neuron* 105:700–712.
- MacDonald AW, Cohen JD, Stenger VA, Carter CS (2000) Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science* 288:1835–1838.
- Masset P, Ott T, Lak A, Hirokawa J, Kepecs A (2020) Behavior- and modality-general representation of confidence in orbitofrontal cortex. *Cell* 182:112–115.
- McCurdy LY, Maniscalco B, Metcalfe J, Liu KY, de Lange FP, Lau H (2013) Anatomical coupling between distinct metacognitive systems for memory and visual perception. *J Neurosci* 33:1897–1906.
- Mei N, Rankine S, Olafsson E, Soto D (2020) Similar history biases for distinct prospective decisions of self-performance. *Sci Rep* 10:5854.
- Metcalfe J, Finn B (2008) Evidence that judgments of learning are causally related to study choice. *Psychon Bull Rev* 15:174–179.
- Miyamoto K, Setsuie R, Osada T, Miyashita Y (2018) Reversible silencing of the frontopolar cortex selectively impairs metacognitive judgment on non-experience in primates. *Neuron* 97:980–989.e6.
- Morales J, Lau H, Fleming SM (2018) Domain-general and domain-specific patterns of activity supporting metacognition in human prefrontal cortex. *J Neurosci* 38:3534–3546.
- Morey RD (2008) Confidence intervals from normalized data: a correction to Cousineau (2005). *TQMP* 4:61–64.
- Moritz S, Woodward TS (2007) Metacognitive training in schizophrenia: from basic research to knowledge translation and intervention. *Curr Opin Psychiatry* 20:619–625.
- Nelson TO, Narens L (1990) Metamemory: a theoretical framework and new findings. In: *The psychology of learning and motivation: advances in research and theory* (Bower GH, ed), pp 125–173. San Diego: Academic.
- Notebaert W, Houtman F, Opstal F, Van Gevers W, Fias W, Verguts T (2009) Post-error slowing: an orienting account. *Cognition* 111:275–279.
- Odegaard B, Grimaldi P, Cho SH, Peters MA, Lau H, Basso MA (2018) Superior colliculus neuronal ensemble activity signals optimal rather than subjective confidence. *Proc Natl Acad Sci USA* 115:E1588–E1597.
- Qiu L, Su J, Ni Y, Bai Y, Zhang X, Li X, Wan X (2018) The neural system of metacognition accompanying decision-making in the prefrontal cortex. *PLoS Biol* 16:e2004037.
- Rabbitt PM (1966) Errors and error correction in choice-response tasks. *J Exp Psychol* 71:264–272.
- Redshaw J, Vandersee J, Bulley A, Gilbert SJ (2018) Development of children's use of external reminders for hard-to-remember intentions. *Child Dev* 89:2099–2108.
- Ridderinkhof KR, Ullsperger M, Crone EA, Nieuwenhuis S (2004) The role of the medial frontal cortex in cognitive control. *Science* 306:443–447.
- Risko EF, Gilbert SJ (2016) Cognitive offloading. *Trends Cogn Sci* 20:676–688.
- Rouault M, McWilliams A, Allen MG, Fleming SM (2018) Human metacognition across domains: insights from individual differences and neuroimaging. *Pers Neurosci* 1:e17.
- Rounis E, Maniscalco B, Rothwell J, Passingham R, Lau H (2010) Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. *Cogn Neurosci* 1:165–175.
- Schulz L, Fleming SM, Dayan P (2021) Metacognitive computations for information search: confidence in control. *BioRxiv*.
- Seow TX, Rouault M, Gillan CM, Fleming SM (2021) How local and global metacognition shape mental health. *Biol Psychiatry*. 90:436–446.
- Shea N, Boldt A, Bang D, Yeung N, Heyes C, Frith CD (2014) Supra-personal cognitive control and metacognition. *Trends Cogn Sci* 18:186–193.
- Shekhar M, Rahnev D (2018) Distinguishing the roles of dorsolateral and anterior PFC in visual metacognition. *J Neurosci* 38:5078–5087.
- Shenhav A, Botvinick MM, Cohen JD (2013) The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron* 79:217–240.
- Shimamura AP (2000) The role of the prefrontal cortex in dynamic filtering. *Psychobiology* 28:207–218.
- Son LK, Schwartz BL (2009) The relation between metacognitive monitoring and control. *Applied Metacognition* (pp. 15–38). Cambridge University Press. <https://doi.org/10.1017/CBO9780511489976.003>.
- Vaccaro AG, Fleming SM (2018) Thinking about thinking: a coordinate-based meta-analysis of neuroimaging studies of metacognitive judgments. *Brain Neurosci Adv* 2:2398212818810591.
- Vaghi MM, Luyckx F, Sule A, Fineberg NA, Robbins TW, De Martino B (2017) Compulsivity reveals a novel dissociation between action and confidence. *Neuron* 96:348–347.
- Wokke ME, Achoui D, Cleeremans A (2020) Action information contributes to metacognitive decision-making. *Sci Rep* 10:3632.
- Ye Q, Zou F, Lau H, Hu Y, Kwok SC (2018) Causal evidence for mnemonic metacognition in human precuneus. *J Neurosci* 38:6379–6387.
- Yeung N, Summerfield C (2012) Metacognition in human decision-making: confidence and error monitoring. *Philos Trans R Soc Lond B Biol Sci* 367:1310–1321.
- Yeung N, Summerfield C (2014) Shared mechanisms for confidence judgments and error detection in human decision making. In: *The cognitive neuroscience of metacognition* (Fleming SM, Frith CD, eds), pp 147–167. Berlin: Springer.
- Yeung N, Botvinick MM, Cohen JD (2004) The neural basis of error detection: conflict monitoring and the error-related negativity. *Psychol Rev* 111:931–959.
- 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.