

Research Articles: Behavioral/Cognitive

Partially overlapping neural correlates of metacognitive monitoring and metacognitive control

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

Cite as: J. Neurosci 2022; 10.1523/JNEUROSCI.1326-21.2022

Received: 25 June 2021

Revised: 10 January 2022

Accepted: 12 January 2022

This Early Release article has been peer-reviewed and accepted, but has not been through the composition and copyediting processes. The final version may differ slightly in style or formatting and will contain links to any extended data.

Alerts: Sign up at www.jneurosci.org/alerts to receive customized email alerts when the fully formatted version of this article is published.

1 **Regular Manuscript**

2

3 **Partially overlapping neural correlates of metacognitive monitoring and metacognitive**
4 **control**

5 **Abbreviated title:** Metacognitive monitoring and control

6

7 Annika Boldt¹ & Sam J Gilbert¹

8 ¹Institute of Cognitive Neuroscience, University College London, Alexandra House, 17-19
9 Queen Square, London WC1N 3AZ, UK

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

11

12 49 pages; 6 figures; 5 tables

13 Word counts: Abstract 150; Introduction 1173; Discussion 1652

14 **Conflict of interests statement:** The authors declare that they have no competing financial
15 or non-financial interests.

16 **Acknowledgements:** This research was funded by the Wellcome Trust, who awarded a Sir
17 Henry Wellcome Postdoctoral Fellowship (206480/Z/17/Z) to AB, and the Economic &
18 Social Research Council (ESRC), who awarded a Research Grant (ES/N018621/1) to SJG.
19 Neither of these funding bodies played a role in the conceptualization, design, data collection,
20 analysis, decision to publish, or preparation of the manuscript. For the purpose of Open
21 Access, the author has applied a CC BY public copyright licence to any Author Accepted
22 Manuscript version arising from this submission. The authors would like to thank the
23 MetaOffloading lab for help with the MRI data collection, Pei-Chun Tsai for help with
24 anatomical labelling, and Carsten Allefeld and Kai Gørgen for useful discussions regarding
25 the data analysis.

26 **Abstract**

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

40

41 **Significance Statement**

42 Models of metacognition commonly distinguish between monitoring (how
43 metacognition is formed) and control (how metacognition is used for behavioural regulation).
44 Research into these facets of metacognition has often happened in isolation. Here, we provide
45 a study which directly investigates the mapping between metacognitive monitoring and
46 metacognitive control at a neural level. We applied multivariate pattern analysis to fMRI data
47 from a novel task in which participants separately rated their confidence (metacognitive
48 monitoring) and how much they would like to use a reminder (metacognitive control). We
49 find support for the notion that the two aspects of metacognition overlap partially but not

50 fully. We argue that future research should focus on how different metacognitive signals are
51 selected for control.

52 **Introduction**

53 Our brains possess a remarkable ability to monitor performance and to then use
54 *metacognition* to control future behavior. For example, if you have low confidence that you
55 will remember a delayed intention (*metacognitive monitoring*; *MetaM*) like regular
56 medication intake, you might set a reminder on your phone (*metacognitive control*; *MetaC*).
57 This distinction between monitoring and control is found in the seminal metamemory
58 framework by Nelson and Narens (1990; see also Flavell, 1976; Kluwe, 1982; Brown, 1987;
59 Efklides, 2008; Shea et al., 2014; Yeung et al., 2004; Fleming & Daw, 2017; Fletcher &
60 Carruthers, 2012; Fleur, Bredeweg, & van den Bos, 2021), which proposes that cognition
61 functions at two distinct levels: the object and the meta level (Figure 1A). Information at the
62 object level about decisions, memories, attention, action and so forth is re-represented at the
63 meta level via a process of *MetaM*. Meanwhile, information at the meta level controls
64 processing at the object level (*MetaC*). Shimamura's (2000) dynamic filtering theory extends
65 the framework by Nelson & Narens (1990), ascribing the role of the object level to posterior
66 cortical regions and the role of the meta level to prefrontal cortex (PFC). The information
67 flow between these regions forms the basis of *MetaM* and *MetaC*.

68 We are only slowly beginning to understand the neural mapping between *MetaM* and
69 *MetaC*. This mapping or link describes the relationship that exists between *MetaM* and
70 *MetaC* on a functional level – are these labels describing the identical process or two
71 different computations with different inputs? This question is important because one rationale
72 for studying *MetaM* is that it can provide insight into *MetaC* (e.g. Boldt & Yeung, 2015;
73 Wokke et al., 2020; Masset et al., 2020; Gherman & Philiastides, 2018; Miyamoto et al.,
74 2018; Odegaard et al., 2018; Bang & Fleming, 2018; Ye et al., 2018; Shekhar & Rahnev,
75 2018). This would be strengthened if the mapping between the two were better understood.
76 Furthermore, dissociations have been found between *MetaM* and *MetaC*. For example, in

77 some circumstances, young children (Redshaw et al., 2018), OCD patients (Vaghi et al.,
78 2017) older adults (Dunlosky & Connor, 1997), and individuals with Autism Spectrum
79 Conditions (Grainger et al., 2016) have a diminished mapping between MetaM and MetaC,
80 which could lead to suboptimal behavioral regulation. However, the potential neural
81 substrates for this variability are unknown.

82 One of the reasons why the MetaM-MetaC mapping has received little attention is
83 that the two aspects of metacognition are usually studied in isolation (though see Koriat et al.,
84 2006, 2014; Mei et al., 2020; Son & Schwartz, 2009; Schulz, Fleming & Dayan, 2021; Qiu et
85 al., 2018). Studies on MetaM commonly explore the variables that affect how confident
86 people feel and the associated neural correlates. For example, neuroimaging studies have
87 identified a widespread network of involved regions, including the rostrolateral prefrontal
88 cortex (rLPFC; Yokoyama et al., 2010; Fleming, Huijgen, & Dolan, 2012; Allen et al., 2017)
89 and also the precuneus specifically for metamemory studies (e.g. McCurdy et al., 2013;
90 Baird, Smallwood, Gorgolewski, & Margulies, 2013; Ye et al., 2018). Moreover, machine-
91 learning techniques have been used to “decode” brain patterns associated with low versus
92 high confidence, using both fMRI (Hebart et al., 2014; Cortese et al., 2016; Morales, Lau &
93 Fleming, 2018) and EEG (Boldt & Yeung, 2015). Research on MetaC, on the other hand, has
94 focused on situations in which metacognitive experiences are utilized for learning,
95 communication, or speed-accuracy tradeoff, to name a few (e.g. Metcalfe & Finn, 2008;
96 Guggenmos et al., 2016; Lak et al., 2020; Shea et al., 2014; Bahrami et al., 2010; Desender et
97 al., 2019; Frömer, Nassar, Bruckner, Stürmer, Sommer, & Yeung, 2021).

98 Most of what we know about the link between monitoring and control comes from the
99 field of cognitive control and error monitoring. Electrophysiological correlates have been
100 found that signal not only when an error has been committed but are also sensitive to correct-
101 trial performance fluctuations (Allain et al., 2004; Yeung, Botvinick, & Cohen, 2004). Such

102 monitoring of errors often results in lower response speed immediately after a mistake, a
103 robust and often-replicated phenomenon termed post-error slowing (Rabbitt, 1966;
104 Danielmeier & Ullsperger, 2011; Notebaert et al., 2009). In addition to errors, conflict signals
105 appear to be monitored by the posterior medial frontal cortex (pmFC) including the dorsal
106 anterior cingulate cortex (dACC). The lateral prefrontal cortex (lpFC) is thought to receive
107 this input and implement cognitive control (Ridderinkhof, Ullsperger, Crone, Nieuwenhuis,
108 2004). It should be noted that participants are often not aware of such errors or response
109 conflicts and that these studies are not directly measuring metacognitive signals.
110 Nevertheless, evidence from this domain suggests that similar brain regions support
111 metacognitive monitoring and control. Qiu and colleagues (2018) conducted four elegant
112 fMRI experiments, using a decision-redecision paradigm: Participants were presented twice
113 in a row with each stimulus and rated both their response and confidence for each
114 presentation. They reasoned that participants would engage metacognitive monitoring for
115 their initial response and use metacognitive control to revise and improve decisions in the
116 redecision phase. Their analyses revealed an involvement of dACC in the first response and
117 lpFC in the second. However, because the order of the decision-redecision phases was always
118 the same, it is impossible to conclude whether the redecision phase really triggered more
119 MetaC or whether the signal observed in lpFC was instead a ‘late’ monitoring one. Another
120 open question is whether MetaM and MetaC rely on similar representations.

121 In order to address these questions, it is necessary to study both aspects of
122 metacognition in a single paradigm, which we did using a cognitive offloading task.
123 Cognitive offloading is the use of physical action to reduce cognitive demand, e.g. setting
124 external reminders rather than relying on internal memory. Previous research has
125 demonstrated a MetaM-MetaC link whereby individuals are more likely to set reminders
126 (MetaC) when they have low confidence in their memory abilities over and above the

127 influence of their actual memory performance (MetaM; Risko & Gilbert, 2016; Hu et al.,
128 2019; Dunn & Risko, 2016). This finding is a robust pattern that can even be observed when
129 reminder setting is not explicitly instructed (Boldt & Gilbert, 2019) or when confidence was
130 measured in an unrelated perceptual task (Gilbert, 2015). Here, we use a decoding approach
131 to examine this link at a neural level.

132 Participants performed a delayed intention task where in separate blocks they engaged
133 in MetaM (how confident am I that I will remember?) or MetaC (how much would I like a
134 reminder?). This allowed us to answer two questions: 1) Do similar brain patterns
135 characterize MetaM and MetaC? If so, 2) Can the neural patterns that characterize specific
136 acts of MetaC be exhaustively characterized in terms of their associated processes of MetaM?
137 We answered these questions by examining cross-classification between MetaM and MetaC:
138 the extent to which a classifier trained on one judgement can decode the other. Insofar as this
139 is possible, this implies a shared neural code for MetaM and MetaC. But if cross-
140 classification is weaker than decoding MetaM and MetaC individually, this implies that their
141 neural bases do not overlap fully.

142

143 **Materials and Methods**

144 *Participants*

145 We trained 29 participants in a behavioral task during a first session. After reviewing
146 their training data, 22 participants returned to the lab for a second MRI session 1 to 21 days
147 later, excluding 7 participants (2 unsuited for MRI due to safety regulations, 2 had extreme
148 staircase values, 3 were unavailable for a second session). Another participant was excluded
149 after scanning due to excessive movement in the scanner. This resulted in a final sample of
150 21 participants, out of which 15 were female and 6 were male. While we determined our
151 sample size based on practical constraints and on available resources, the final sample size of

152 $N=21$ is nevertheless in accordance with previous MRI studies using similar methods
153 (Morales et al., 2018, Qiu et al., 2018; Hebart et al., 2014). Participants were 20.3 years on
154 average (18 – 26 years and paid £36 for their participation in both sessions (about 90 and 150
155 minutes). All participants were right-handed, had intact color vision, no uncorrected visual
156 impairments and had not been diagnosed with any psychiatric or neurological disorders. All
157 testing was approved by the local ethics committee and participants gave informed consent
158 prior to taking part in the study.

159

160 *Experimental Design*

161 In order to investigate the extent to which neural patterns associated with MetaM and
162 MetaC are similar or distinct we had to study both aspects of metacognition within a single
163 paradigm. Participants underwent short miniblocks of ongoing shape discrimination trials.
164 For this ongoing task, participants had to quickly and accurately decide whether an array of
165 colored shapes grouped around a fixation dot looked on average more like a circle or a square
166 (De Gardelle & Summerfield, 2011) by pressing one of two buttons. The response categories
167 were equally likely. During some of these miniblocks, participants also had to maintain a
168 delayed intention to press a different button if the stimulus appeared in a target color (Figure
169 1B). Participants were allowed to use reminders (cognitive offloading) to support their
170 prospective memory in approximately half of the miniblocks, which meant that the central
171 fixation dot of the stimulus took on the target color for the duration of the miniblock. Instead
172 of having to rely on their memory, participants could then simply wait for the color of the
173 shapes to match the color of the fixation dot, making the fulfilment of the delayed intention
174 much easier. There were 12 colors, placed equidistant in RGB space. Within each miniblock,
175 colors were drawn without replacement. There was only one target color per miniblock,

176 presented at the beginning of the miniblock, and its occurrence during the ongoing-task trials
177 always terminated the miniblock.

178 The task comprised three within-subject experimental conditions (20% Baseline, 40%
179 MetaM, and 40% MetaC; see Table 1) each structured into miniblocks. A miniblock
180 comprised presentation of a target color (except for in the Baseline condition which had no
181 prospective memory element), a single metacognitive rating or cursor placement, followed by
182 3-7 ongoing-task trials. The number of trials per miniblock was drawn from an exponential
183 distribution with a mean of $\mu = 1.1$, in other words shorter miniblocks were more frequent
184 than longer miniblocks. Each of the eight blocks consisted of 94 shape trials spread unevenly
185 across 40 miniblocks (see Figure 1C). The critical difference between our two key conditions
186 was the metacognitive rating given about the target color before each miniblock. In the
187 MetaC condition, participants reported how much they would want to set a reminder to help
188 them remember this target color. The higher the rating given by the participant, the greater
189 the likelihood of receiving a reminder, which occurred on approximately 50% of miniblocks.
190 More specifically, ratings larger than the moving median of the past 20 MetaC ratings were
191 assigned a reminder, whereas ratings below this cutoff had to be solved using only unaided
192 memory. In the MetaM condition, participants reported their prospective confidence in
193 remembering the target color. However, this had no influence on the likelihood of receiving a
194 reminder, which occurred on a randomly selected 50% of miniblocks. In other words, the two
195 conditions also differed in the relationship between participants' ratings and the provision of
196 reminders. In the MetaM condition, participants' ratings had no influence on whether or not
197 they received a reminder. In the MetaC condition, on the other hand, which miniblock
198 contained a reminder was largely determined by participant's ratings. Therefore, in the
199 MetaM condition participants engaged in metacognitive monitoring but did not exercise
200 metacognitive control. In the MetaC condition they exercised control to make a decision

201 which is known to be guided by metacognition (Boldt & Gilbert, 2019; Gilbert, 2015).

202 However, they were not explicitly asked to make a direct metacognitive judgment.

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

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

222 The study comprised two sessions. The purpose of the first session was assessment of
223 MRI safety, completion of a pre-study questionnaire on how much participants liked the 12
224 colors used in the task, and training in the behavioral task (presented in MATLAB using
225 Psychtoolbox3; Kleiner, Brainard, Pelli, Ingling, Murray, & Broussard, 2007). Participants

226 first completed eight practice blocks, each introducing them to a new aspect of the paradigm.
227 They then completed four experimental runs that were identical to the task they would have
228 to complete whilst in the scanner, each lasting ~9 minutes. During the second session,
229 participants first underwent two practice blocks outside of the scanner (each lasting ~5
230 minutes) to remind them of the task before they completed eight runs in the scanner, with a 6-
231 minute T1 scan between the fourth and fifth run. One participant only completed six blocks
232 due to feeling unwell inside the scanner. Due to the unbalanced design, we decided to
233 exclude this participant from all multivariate analyses.

234 At the end of the second session, participants were furthermore asked to fill in a post-
235 experiment questionnaire, asking them to rate the liking of all colors again, together with how
236 difficult they found them and several additional questions to determine whether they
237 perceived the MetaM and MetaC conditions as similar, how much control they felt during
238 these conditions, how they used the reminders depending on whether or not they asked for
239 them, and how they approached each rating. The orientation of the rating scales was flipped
240 halfway through the experiment to avoid confounding visuomotor processes with low versus
241 high ratings. The order of scale orientations, response keys for the shape task, and the order
242 of the conditions were counterbalanced across participants.

243

244 *MRI Data Collection and Preprocessing*

245 We used a 1.5T Siemens Avanto scanner with a 32-channel head coil and MRI-safe
246 button boxes. We acquired both T1-weighted structural images, as well as T2*-weighted
247 echoplanar images (EPI; 64 x 64; 3.2x3.2x3.2 mm voxels) with blood oxygen level-
248 dependent (BOLD) contrast. We used a multiband acquisition sequence with acceleration
249 factor = 3, TE = 54.8 ms, flip angle = 75°, to record 39 interleaved, axial slices (3.2mm thick,
250 oriented approximately to the anterior commissure - posterior commissure plane). This

251 allowed us to cover most of the brain with an effective repetition time of 1.3s per volume.
252 Encoding phase direction was anterior to posterior. Functional scans were acquired in eight
253 runs, each comprising 410 volumes (~9 min). The first five volumes in each session were
254 discarded to allow for T1 equilibration effects. Between the fourth and fifth functional scans,
255 an approximately 6 min T1-weighted MPRAGE structural scan was collected.

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

263

264 *Statistical Analysis*

265 Analyses of behavioral data were conducted using R version 3.6.0 (“Planting of a
266 Tree”) with the additional packages *plyr*, *plotrix*, *Hmisc*, *R.matlab*, *viridis*, *effsize*,
267 *raincloudplots*, *ggplot2*, *grid*, *gridExtra*, and *Rmisc*. Statistical tests were conducted two-
268 sided if not stated otherwise. For *t*-tests we reported effect sizes as Cohen’s *d*, and for
269 ANOVAs as partial eta square η_p^2 . For the fMRI analyses, the volumes acquired during the
270 eight sessions were treated as separate time series. For each time series, the variance in the
271 BOLD signal was decomposed with a set of regressors in a general linear model. Three
272 regressors were generated to code for the target color presentation and the rating as a 7s
273 boxcar, separately for miniblock and rating conditions (Baseline in MetaM blocks, low
274 MetaM rating, high MetaM rating in MetaM blocks and Baseline in MetaC blocks, low
275 MetaC rating, high MetaC rating in MetaC blocks). Six additional regressors were generated

276 that represented effects of no interest, specifically, stimulus presentation as a stick function,
277 separately for targets and non-targets, the ongoing task spanning from the onset of the first to
278 the last shape stimulus of the miniblock, separately for whether there was a prospective-
279 memory requirement (Baseline vs. MetaM and MetaC) and the time when the computer
280 revealed to the participant whether they were allowed to use a reminder as a stick function,
281 separately for Reminder and Own Memory miniblocks. All regressors were convolved with a
282 canonical hemodynamic response function. The regressors outlined above, along with six
283 regressors representing residual movement-related artefacts and the mean over scans
284 comprised the full model for each session. The data and model were high-pass filtered at a
285 cutoff of 1/128 Hz. Parameter estimates for each regressor were calculated from the least
286 mean squares fit of the model to the data. Effects of interest were assessed in a random-effect
287 analysis by first forming subject-specific contrasts subtracting the Baseline from the other
288 two conditions. The resulting contrast images were entered into a repeated-measures
289 ANOVA using nonsphericity correction (Friston, Glaser, Henson, Kiebel, Phillips, &
290 Ashburner, 2002), representing a condition agnostic selection contrast to identify a network
291 of regions active in the rating task. Results are reported applying a height threshold of $p <$
292 0.001 uncorrected in conjunction with an extent threshold determined by SPM12 to achieve p
293 < 0.05 familywise error correction for multiple comparisons across the whole brain volume.
294 Region of interest (ROI) analyses were conducted by extracting subject-specific contrast
295 estimates from the resulting ROIs with the toolbox MarsBaR (Brett, Anton, Valabregue,
296 Poline, 2002), then entering the resulting data into an ANOVA in R using the same correction
297 procedure described above.

298 The logic behind the key analysis of our study was the following: Replicating and
299 extending previous findings (Hebart et al., 2014; Cortese et al, 2016; Morales et al., 2018) we
300 first trained separate classifiers to detect A) whether participants were in a high or low

301 confidence state (MetaM), and B) whether they had high or low desire for a reminder
302 (MetaC). These classifiers could then also be combined in a cross-classification analysis, that
303 is whether a classifier trained on MetaM ratings can also predict MetaC ratings (and vice
304 versa). Insofar as this cross-classification is possible, this suggests shared brain
305 representations for both aspects of metacognition. Going one step further, we then compared
306 within-category classification to cross-classification accuracy to distinguish between two
307 possible patterns of results: If MetaM and MetaC are based on the exact same
308 representational code, there should be no difference in classification accuracy. If, on the other
309 hand, MetaM and MetaC share partially overlapping patterns, we should find significantly
310 higher classification accuracy for within- than across-category classification, but
311 significantly-different-from-zero accuracy for cross-classification.

312 For the multivariate-pattern analyses, we used The Decoding Toolbox (TDT; Hebart,
313 Gørgen, & Haynes, 2015), based on the beta images resulting from the previously described
314 general linear models (except that the models were re-fit to unsmoothed, unnormalized data
315 and the MetaM and MetaC boxcar regressors were split into two regressors each using a
316 median split on the respective metacognitive rating). When we ran our four separate decoding
317 analyses, two drew the training and testing data from the same condition (low vs. high ratings
318 for the MetaM and MetaC conditions respectively; defined by block-, condition- and subject-
319 wise median splits), whereas the other two cross-classified (train on low vs. high MetaM
320 ratings and test on high vs. low MetaC ratings and vice versa; note that the rating scale had to
321 be flipped for MetaC as low confidence implies high desire for a reminder). For each of these
322 analyses, a linear support vector machine (SVM) was trained to discriminate between low
323 versus high ratings given the patterns of BOLD activity across voxels. Given the alternating
324 block design and the fact that the orientation of the scale was flipped halfway through the

325 study, we had two low and two high rating images available for each training or testing fold,
326 resulting in a 2-fold procedure (see Figure 1D).

327 We used a whole-brain searchlight approach (Kriegeskorte, Goebel, & Bandettini,
328 2006), meaning that for each voxel a separate SVM was built, fitted to the beta values within
329 a sphere with a radius of 3 voxels (9.6 mm). This resulted in three-dimension decoding
330 accuracy maps in native space for each participant and analyses. Decoding accuracy is
331 calculated relative to chance level (subtracted by 50%, so a 5% accuracy corresponds to
332 55%). These maps were then normalized into Montreal Neurological Institute (MNI) space
333 (using the same normalization parameters as the univariate analyses) and smoothed using a
334 Gaussian kernel (full-width half-maximum, 4mm). Please note that this kernel was half of the
335 one used for the univariate analyses. This was done to avoid excessive smoothing, given that
336 the searchlight analysis already imposes spatial smoothing on the data. The resulting images
337 were entered into a one-sample *t*-test using SPM12. This allowed assessment of voxels
338 showing consistently higher decoding accuracy in a random-effect analysis. We note that the
339 suitability of second-level *t*-tests has been challenged for information-like measures such as
340 classification accuracy, where classifier performance can meaningfully be above, but not
341 below, chance levels (Allefeld, Gørgen & Haynes, 2016; Hirose, 2020). However, this
342 characteristic does not apply to our two key hypothesis-testing analyses. For the cross-
343 classification between MetaM and MetaC, high MetaM could either predict higher or lower
344 MetaC. For the comparison between within- and cross-classification, accuracy for one
345 classification could be higher or lower than the other. Therefore, in both cases our statistical
346 tests are valid because they are performed on data that could meaningfully take values both
347 above and below zero.

348 Along with the main MVPA analyses described above, we conducted an additional
349 analysis. Here, we used a similar approach to the univariate ROI analysis described above by

350 defining a condition agnostic contrast (the mean of all four decoding analyses), extracting
351 ROIs with significantly above-chance decoding accuracy and then entering the resulting
352 classification accuracies into a repeated-measures ANOVA with factors ROI, Training
353 condition (MetaM/MetaC), and Classification type (within-condition/cross-classification).

354

355 **Results**

356 *Behavioral Results*

357 Our sample included 22 participants, one of which was excluded due to excessive
358 motion in the MRI scanner, see Methods for more details. Participants performed the tasks
359 with a high level of accuracy (mean shape-discrimination accuracy = 93.4%, SEM = 0.84%;
360 non-significant shape bias, $t(20) = 1.2$, $p = 0.25$, $d = 0.26$; mean target-detection rate =
361 88.1%, SEM = 3.37%; NB chance target-detection accuracy would be 8.3%; false alarm rate
362 = 0.8%). With our design, we decided against using a direct manipulation of difficulty (such
363 as spacing some colors closer to each other in color space) as this would have made it
364 difficult to interpret any effect of confidence due to its inherent confound with a difficulty
365 manipulation. Instead, we relied upon natural fluctuations in confidence, caused for example
366 by individual preferences for colors or fatigue. Figure 2A shows that average, unaided
367 memory performance varied across colors with some colors (e.g. the 4th color, a shade of
368 green) being associated with lower accuracy when participants had to remember this target
369 color unaided by a reminder. Moreover, this figure shows that not all participants had the
370 same inherent color-difficulty profile and that instead some participants perceived particular
371 colors as more difficult than others. Performance in the Baseline condition was high. Here, an
372 indicator of compliance with instructions is participants' placement of the cursor between
373 two thin lines marked on the scale (Figure 2B). Participants reported that those lines were
374 difficult to see in the scanning session. Nevertheless, their cursor locations peaked around the

375 marked location and landed within the marked positions on approximately half the trials, $\delta =$
376 47.6%.

377 We next established that the reminders aided participants in their fulfilment of the
378 delayed intentions by comparing target-detection error rates for miniblocks in which
379 participants had to use their own memory (fixation dot stayed white) to miniblocks in which
380 they were allowed to use a reminder (fixation dot took on target color), shown in Figure 2C.
381 In both conditions, error rates were reduced when reminders could be used ($F(1,20) = 20.5$, p
382 < 0.001 , $\eta^2_p = 0.51$; $t_s > 3.4$, $p_s < 0.01$, $d_s > 0.37$ when tested separately for the MetaM and
383 MetaC conditions). Error rates did not differ significantly between conditions, $F < 1$, nor was
384 there an interaction between the two factors, $F < 1$.

385 When asked explicitly after the experiment how similar they perceived the two
386 conditions, participants rated the conditions as similar but not identical ($M = 0.68$ on a scale
387 from 0 = 'totally different' to 1 = 'exactly the same'; min = 0.28; max = 0.98). In fact, we
388 found that participants' perception of the two conditions differed in how much control
389 participants felt they had over the reminders. On a scale ranging from 0 = no control to 1 =
390 full control, participants rated the MetaM condition with a mean of $M = 0.32$ (min = 0.00;
391 max = 0.88) and the MetaC condition with a mean of $M = 0.80$ (min = 0.06; max = 0.98).
392 This difference was significant, $t(20) = 6.4$; $p < 0.001$, $d = 1.94$. This shows that participants
393 were able to grasp the key difference that distinguished the two conditions.

394 We furthermore aimed to rule out that any condition differences found in the pattern
395 classification analyses could be caused by behavioral differences in how the different ratings
396 were approached. Firstly, Figure 3A shows that the average ratings participants gave for each
397 individual color were almost indistinguishable whether they were giving a metacognitive-
398 monitoring or metacognitive-control rating. In fact, if we correlated the average ratings for
399 each color for each individual participant, there was an average relationship of $r = 0.76$ with

400 19 out of 21 participants showing a significant, positive relationship between the MetaM and
401 the MetaC rating for different colors. Relatedly, participants' rating and rating RT
402 distributions for the two types of ratings were closely matched (Figures 3B and 3C). It is
403 important to note that participants did not receive any instructions to use these scales in the
404 same way (except for being asked to use the entire range of the scale in both cases).

405 Furthermore, neither of the metacognitive rating conditions showed a systematic
406 relationship between confidence and accuracy: For retrospective confidence judgements, it is
407 commonly found that these correlate, that is participants express lower confidence on errors
408 than on correct trials (confidence resolution or type-II sensitivity). In the MetaC condition on
409 the other hand, participants' ratings triggered reminders, so we would expect to see the
410 opposite pattern: Trials for which they expressed a high need for a reminder should naturally
411 be the ones on which they were allowed to offload and error rates should therefore be lower.
412 However, we found no significant difference between correct- and error-trial ratings in any of
413 the four conditions (MetaC reminder, $t(14) = 0.2$, $p = 0.88$, $d = 0.04$; MetaC own memory,
414 $t(19) = 0.1$, $p = 0.95$, $d = 0.02$; MetaM reminder, $t(17) = 1.0$, $p = 0.34$, $d = 0.30$; MetaM own
415 memory, $t(20) = 1.1$, $p = 0.28$, $d = 0.23$; participants with missing data excluded from the
416 respective analysis). We furthermore correlated the dichotomous accuracy vector with our
417 continuous confidence measure for all four data cells, separately for each participant. The
418 distributions of these correlations are shown in the right panels of Figure 3D. None were
419 significantly different from zero, $ts < 1.0$, $ps > 0.32$. Taken together, both the prospective
420 nature of the ratings in the present task (i.e. participants might have felt they needed to invest
421 more into trials in which they felt less confident or wanted a reminder more) and our unique
422 offloading design could potentially have led to a reduced confidence resolution, but this was
423 the case for both rating conditions.

424

425 *Univariate fMRI Results*

426 We first performed univariate analyses to identify brain regions activated by the
427 requirement to encode new intentions and make metacognitive judgements about them. We
428 therefore averaged across the two metacognition conditions (MetaM and MetaC) and
429 compared them to the Baseline condition, allowing us to find regions of interest (ROIs)
430 activated by our task. After family-wise error correcting for multiple comparisons, this
431 contrast revealed seven regions showing increased BOLD signal in the metacognitive
432 conditions (see Table 2 and Figures 4A and 4B).

433 Within the seven ROIs, activity was then compared between the metacognition
434 conditions. More specifically, activity was extracted in two separate contrasts (MetaM >
435 Baseline and MetaC > Baseline) and then compared. Note that this comparison is orthogonal
436 to the initial selection contrast and therefore unbiased (Kriegeskorte, Simmons, Bellgowan, &
437 Baker, 2009). BOLD signal was higher for the MetaC than the MetaM condition in all seven
438 ROIs (Figure 4C) and this main effect was significant when examined in a ROI (7) x
439 Condition (2: MetaC/MetaM) repeated-measures ANOVA, $F(1,20) = 8.1, p = 0.01, \eta_p^2 =$
440 0.29 . There was furthermore a reliable main effect of ROI, $F(7,140) = 7.8, p < 0.001, \eta_p^2 =$
441 0.28 , as well as a significant interaction of the two factors, $F(7,140) = 3.4, p < 0.01, \eta_p^2 =$
442 0.14 , reflecting that the absolute signal change and also the difference in signal change was
443 larger in some ROIs compared to others. Taken together, these results show that regions
444 which respond to the conditions requiring delayed intentions and metacognitive judgments
445 showed higher activity when participants rated how much they would like a reminder
446 (MetaC) compared to how confident they were (MetaM).

447 We repeated the univariate analyses for deactivations, revealing six “task-negative”
448 regions showing decreased signal in the conditions requiring delayed intentions and
449 metacognitive judgments compared with baseline (Figure 5 and Table 3). These regions

450 included the cingulate and paracingulate cortices, supplementary motor area, supramarginal
451 gyrus, middle and inferior temporal gyri, occipital gyri, and anterior cingulate gyrus. Within
452 these task-negative ROIs, there was more deactivation when participants rated how confident
453 they were (MetaM) compared to how much they would like a reminder (MetaC), however,
454 BOLD signal did not differ significantly between the MetaC and the MetaM condition,
455 $F(1,20) = 1.3$, $p = 0.26$, $\eta_p^2 = 0.06$. There was a reliable main effect of ROI, $F(5,100) = 18.2$,
456 $p < 0.001$, $\eta_p^2 = 0.48$. The interaction was not significant, $F < 1$.

457

458 *Multivariate fMRI Results*

459 The multivariate analyses allowed us to address our two key questions: 1) Do the
460 brain patterns of different metacognitive experiences also distinguish different acts of
461 control? and 2) Can the neural patterns that characterize specific acts of metacognitive
462 control be exhaustively characterized in terms of their associated metacognitive experiences?
463 In a first analysis, we attempted to decode confidence (MetaM). Figures 6A and 6B show the
464 resulting decoding accuracy maps corrected for chance level and multiple comparisons,
465 resulting in nine clusters that contained meaningful information when predicting whether the
466 brain was currently in a low or high confidence state including the anterior cingulate gyrus,
467 parietal occipital sulcus, central sulcus, superior parietal lobule, superior occipital gyrus,
468 cuneus, precuneus, supplementary motor area, occipital fusiform gyrus, calcarine cortex;
469 superior corona radiata, and precentral gyrus (Table 4).

470 We then repeated the equivalent analysis for the MetaC condition, again successfully
471 decoding whether participants gave a low or high rating (i.e. desire for a reminder) from five
472 clusters including the occipital pole, lateral occipital cortex, superior parietal lobule, superior
473 frontal gyrus (medial segment), middle temporal gyrus (see Table 4). Together these analyses
474 show that the neuroimaging data contains meaningful patterns that distinguish both different

475 metacognitive experiences (low vs. high confidence) and different acts of metacognitive
476 control (low vs. high desire for a reminder).

477 Having established the existence of meaningful patterns across the brain that
478 distinguish different levels both of MetaM and MetaC, we could then ask whether it was
479 possible to cross-classify the two aspects of metacognition. More specifically, we trained
480 classifiers to distinguish low from high confidence beta images (MetaM) and tested them to
481 predict high versus low MetaC ratings. Note that an inverse relationship is expected between
482 MetaM and MetaC ratings, i.e. low confidence predicts high desire for reminder and vice
483 versa. Therefore one of the scales was inverted in order to perform this analysis. Above-
484 chance classification accuracy can be interpreted as overlapping patterns encoding both
485 MetaM and MetaC. The same analysis was then applied to the opposite direction (train on
486 MetaC, test on MetaM). Importantly, we found overlapping patterns that encode these
487 different types of metacognitive ratings. However, only for the latter analysis direction (train
488 on MetaC, test on MetaM) did we find above-chance classification accuracy after correcting
489 for multiple comparisons. The surviving cluster was located in the left superior and middle
490 frontal gyri. These findings show that brain patterns associated with different metacognitive
491 experiences (low vs. high confidence) also distinguish different acts of metacognitive control
492 (low vs. high desire for a reminder).

493 To address our second key question, we compared classification accuracy resulting
494 from the two different types of classification analyses described above: within-category (test
495 on MetaM and train on MetaM; test on MetaC and train on MetaC) versus across-category
496 classification (i.e. cross-classification: test on MetaM and train on MetaC; test on MetaC and
497 train on MetaM). We first performed a condition-blind analysis by averaging across all four
498 decoding analyses. This identified ROIs that contain information in one or more of the
499 analyses in an unbiased manner, yielding significant effects in the occipital pole, middle

500 occipital gyrus, parietal cortex (superior parietal lobule, precuneus), superior frontal gyrus,
501 middle frontal gyrus; precentral gyrus (see Table 5 and Figures 6C and 6D). Within the
502 resulting ROIs, classification accuracies in the four analyses could then be compared (see
503 Figure 6E) to address the question whether decoding accuracy differed significantly between
504 the within-condition classification and the cross-classification analyses. Taking an analogous
505 approach to our univariate analysis, these comparisons were unbiased because they were
506 orthogonal to the analysis used to define the ROIs. We entered the classification accuracies
507 from these regions into a repeated-measures ANOVA with factors ROI, Training condition
508 (MetaM/MetaC) and Classification type (within-condition/between-condition cross-
509 classification). There was a significant main effect of Classification type, $F(1,19) = 6.2, p =$
510 $0.02, \eta_p^2 = 0.25$, with higher classification accuracy for within-condition classifications than
511 between-condition cross-classifications. This finding can be interpreted as partially
512 overlapping neural representations between MetaM and MetaC as opposed to perfect overlap
513 between the patterns associated with the two aspects of metacognition. Moreover, there was
514 no effect of the conditions on which the classifier was trained or which ROI was analyzed, F_s
515 < 1 . We found a significant interaction between ROI and category (within vs. between
516 classification), $F(6,114) = 2.4, p = 0.03, \eta_p^2 = 0.11$, reflecting that the difference between
517 within-condition and across-condition decoding analyses was larger in some ROIs compared
518 to others. No other interactions were significant, $F_s < 1$. In sum, while our results
519 demonstrate overlapping patterns between metacognitive monitoring and control, they also
520 suggest that patterns of metacognitive control cannot exhaustively be characterized by
521 associated patterns of metacognitive monitoring when participants report their confidence.

522

523 **Discussion**

524 Metacognitive monitoring is only valuable insofar as it can subsequently influence
525 control. And metacognitive control can only occur if there are metacognitive representations
526 to begin with, which can then be utilized to adjust future behaviour. The two processes must
527 therefore be intimately related, yet the mapping between them requires further study,
528 especially on a neural level. Here we report three main findings: 1) we can separately decode
529 metacognitive monitoring and metacognitive control; 2) brain patterns of different levels of
530 metacognition monitoring (low vs. high confidence) also distinguish different acts of
531 metacognitive control (low vs. high desire for a reminder); and 3) this overlap in patterns
532 while significant is only partial. These findings suggest that patterns of brain activity
533 corresponding to specific acts of metacognitive control are partially, but not fully,
534 characterized by associated acts of metacognitive monitoring.

535 Our cross classification analysis revealed involvement of the left superior and middle
536 frontal gyri, which form part of the lateral prefrontal cortex (laPFC) in both metacognitive
537 monitoring and control. The role of the laPFC in metacognition has already been highlighted
538 by previous studies, suggesting a role in domain-general metacognition (Morales et al., 2018;
539 see also Vaccaro & Fleming, 2018), in the readout of sensory information as an input for
540 confidence signals (Shekhar & Rahnev, 2018), and more broadly in a mediating role of more
541 rostral parts of laPFC in metacognitive accuracy (Fleming, Weil, Nagy, Dolan, & Rees, 2010;
542 Rounis, Maniscalco, Rothwell, Passingham, & Lau, 2010). Crucially, the laPFC has also been
543 implied in metacognitive control (Qiu et al, 2018; for reviews see Seow, Rouault, Gillan, &
544 Fleming, 2021; Shimamura, 2000; Fleming & Dolan, 2014) matching its more general
545 proposed involvement in cognitive control (MacDonald, Cohen, Stenger, & Carter, 2000;
546 Ridderinkhof et al., 2004). Our study therefore extends this growing body of research that
547 implies an involvement of the lateral prefrontal cortex in metacognition and cognitive
548 control.

549 Seeing as MetaC could not be characterized exhaustively in terms of the MetaM
550 judgments we investigated, this raises the obvious question of which other signals might
551 contribute to MetaC. We consider two main possibilities. The first possibility is that non-
552 metacognitive signals also play a role in influencing MetaC. A wide variety of signals may be
553 relevant here, such as motivation, the costs and rewards associated with different levels of
554 performance, serial dependencies, fatigue, states of interoceptive and bodily awareness
555 reflecting endogenous signals like arousal (Allen et al., 2016; Hauser et al., 2017; Rouault,
556 McWilliams, Allen, & Fleming, 2018) and so on. This influence of non-metacognitive
557 signals on metacognitive control was already acknowledged in the seminal paper by Nelson
558 and Narens (1990) introducing their metamemory framework. The influence of a wide variety
559 of signals on control is also central to an influential model from the cognitive control
560 literature, the Expected Value of Control model (EVC; Shenhav, Botvinick, & Cohen, 2013).
561 This model emphasizes the flexibility with which different control signals are selected, based
562 on the costs and benefits associated with these signals. The model proposes that the dorsal
563 anterior cingulate cortex integrates both costs and benefits to form the expected value of
564 control. Seeing as MetaC may involve the integration of multiple relevant signals, including
565 the products of MetaM and additional non-metacognitive signals as well, this could
566 potentially explain the greater univariate signal we observed for the MetaC than the MetaM
567 condition. This suggests the incorporation of additional processes into the MetaC judgement
568 beyond those involved in MetaM. We also note that the factor of within- versus cross-
569 classification interacted significantly with region, even though there was no main effect of
570 region. This suggests that the overlap between MetaM and MetaC is greater in some regions
571 than others.

572 A second possible contribution to the MetaC condition is the integration of additional
573 metacognitive signals, beyond the confidence judgement required by the MetaM condition. In

574 our task, for instance, participants' desire for reminder might have been influenced not only
575 by confidence in their prospective memory but also confidence in their perceptual
576 judgements. Consistent with this, behavioral evidence suggests that confidence judgements
577 are influenced by a variety of domain-general and domain-specific signals (Gilbert, 2015;
578 Kantner, Solinger, Grybinas, Dobbins, 2018; Rouault et al., 2018). Confidence can be
579 regarded as an explicit representation of uncertainty, and uncertainty exists at multiple levels
580 throughout the brain (as noted by the Bayesian brain hypothesis; Knill & Pouget, 2004).
581 Therefore, the metacognitive signals measured in the MetaM condition probably form only a
582 subset of the metacognitive signals which may have contributed to MetaC judgements.

583 Our paradigm involved measurement of only a single MetaC judgment, which may
584 have been influenced by multiple MetaM signals. In reality, there are multiple types of both
585 MetaM and MetaC. Take for example the situation of a foreign language student studying for
586 a test at her desk during the early evening hours. The student reads a word on a flashcard and
587 we can assume she has access to two relevant metacognitive signals: On the one hand there is
588 the certainty with which the word is perceived in the waning light, the other is the certainty
589 with which the word is recognized from memory. The former confidence should guide her
590 decision whether or not to switch on her desk lamp. The latter confidence should guide her
591 decision whether or not to place the flashcard on the pile marked as 'restudy'. Similarly, the
592 same confidence signal could lead to opposite consequences depending on the situation as
593 shown by Carlebach & Yeung (2021). The authors report that low confidence leads to advice-
594 seeking when the quality of the advice is known and high. However, when the quality of the
595 advice is unknown, people tend to seek advice especially when they have high confidence to
596 test the accuracy of the advisor. How does the brain then 'harvest' these various confidence
597 signals and route them to the appropriate act(s) of metacognitive control? How does it
598 flexibly switch to a different set of signals when required to do so? How are metacognitive

599 signals weighted by past rewards, and how do such weightings shift when our goals change?
600 Questions such as these could potentially be addressed by adapting the present paradigm to a
601 situation involving two or more forms of metacognitive monitoring and control.

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

622 Our findings bear some interesting parallels to another recent decoding study: Mei
623 and colleagues (2020) reported the results from two behavioral experiments, each focused on

624 a different type of prospective decision (belief of successfully classifying a visual stimulus
625 vs. deciding whether or not to attend to the stimulus during the upcoming trial). The authors
626 found that it was possible to use the data from one experiment (awareness ratings, confidence
627 ratings and accuracy in previous trials) to predict the prospective decision from the respective
628 other experiment and vice versa. This cross-classification analysis therefore highlights
629 similarities of metacognitive monitoring (in this case: beliefs of successfully classifying the
630 upcoming stimulus) and metacognitive control (in this case: decision to attend), showing that
631 both aspects of metacognition appear in the context of the same behavioral precursors.

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

644 In sum, our study delineates the similarities and divisions between neural correlates of
645 metacognitive monitoring and control. Ultimately, understanding the link between
646 monitoring and control could inform interventions such as metacognitive training in
647 conditions including brain injury (Fleming et al., 2017), schizophrenia (Moritz & Woodward,
648 2007) and OCD (Fisher & Wells, 2008). We propose that a full understanding of the

649 relationship between monitoring and control will require a focus on the ways in which
650 distinct metacognitive signals are integrated and selectively routed to appropriate acts of
651 metacognitive control.
652

653 **References**

- 654 Allain, S., Carbonnell, L., Falkenstein, M., Burle, B., & Vidal, F. (2004). The modulation of
655 the Ne-like wave on correct responses foreshadows errors. *Neuroscience Letters*,
656 372(1–2), 161–166. <https://doi.org/10.1016/j.neulet.2004.09.036>
- 657 Allefeld, C., Görden, K., & Haynes, J. D. (2016). Valid population inference for information-
658 based imaging: From the second-level t-test to prevalence inference. *NeuroImage*, 141,
659 378–392. <https://doi.org/10.1016/j.neuroimage.2016.07.040>
- 660 Allen, M., Frank, D., Schwarzkopf, D. S., Fardo, F., Winston, J. S., Hauser, T. U., & Rees, G.
661 (2016). Unexpected arousal modulates the influence of sensory noise on confidence.
662 *ELife*, 5, 1–17. <https://doi.org/10.7554/eLife.18103>
- 663 Allen, M., Glen, J. C., Müllensiefen, D., Schwarzkopf, D. S., Fardo, F., Frank, D., ... Rees,
664 G. (2017). Metacognitive ability correlates with hippocampal and prefrontal
665 microstructure. *NeuroImage*, 149(February), 415–423.
666 <https://doi.org/10.1016/j.neuroimage.2017.02.008>
- 667 Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R., & Kievit, R. A. (2019). Raincloud
668 plots: a multi-platform tool for robust data visualization. *Wellcome Open Research*, 4,
669 63. <https://doi.org/10.12688/wellcomeopenres.15191.1>
- 670 Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010).
671 Optimally interacting minds. *Science*, 329(5995), 1081–1085.
672 <https://doi.org/10.1126/science.1185718>
- 673 Baird, B., Smallwood, J., Gorgolewski, K. J., & Margulies, D. S. (2013). Medial and lateral
674 networks in anterior prefrontal cortex support metacognitive ability for memory and

- 675 perception. *The Journal of Neuroscience : The Official Journal of the Society for*
676 *Neuroscience*, 33(42), 16657–16665. [https://doi.org/10.1523/JNEUROSCI.0786-](https://doi.org/10.1523/JNEUROSCI.0786-13.2013)
677 [13.2013](https://doi.org/10.1523/JNEUROSCI.0786-13.2013)
- 678 Bang, D., & Fleming, S. M. (2018). Distinct encoding of decision confidence in human
679 medial prefrontal cortex. *Proceedings of the National Academy of Sciences*, 115(23),
680 6082–6087. <https://doi.org/10.1073/pnas.1800795115>
- 681 Boldt, A., & Gilbert, S. J. (2019). Confidence guides spontaneous cognitive offloading.
682 *Cognitive Research: Principles and Implications*, 4(1), 45.
683 <https://doi.org/10.1186/s41235-019-0195-y>
- 684 Boldt, A., & Yeung, N. (2015). Shared Neural Markers of Decision Confidence and Error
685 Detection. *Journal of Neuroscience*, 35(8), 3478–3484.
686 <https://doi.org/10.1523/JNEUROSCI.0797-14.2015>
- 687 Brett, M., Anton, J.-L., Valabregue, R., & Poline, J.-B. (2002). Region of interest analysis
688 using an SPM toolbox. In *8th International Conference on Functional Mapping of the*
689 *Human Brain*. Sendai, Japan.
- 690 Brown, A. L. (1987). Metacognition, executive control, self-regulation, and other more
691 mysterious mechanisms. In F. E. Weinert & R. H. Kluwe (Eds.), *Metacognition,*
692 *motivation, and understanding* (pp. 95–116). Hillsdale, NJ: Erlbaum.
- 693 Carlbach, N., & Yeung, N. (2020, October 25). Flexible use of confidence to guide advice
694 requests. *PsyArXiv*. <https://doi.org/10.31234/osf.io/ctyqp>
- 695 Cortese, A., Amano, K., Koizumi, A., Kawato, M., & Lau, H. (2016). Multivoxel
696 neurofeedback selectively modulates confidence without changing perceptual

- 697 performance. *Nature Communications*, 7(1), 13669.
698 <https://doi.org/10.1038/ncomms13669>
- 699 Danielmeier, C., & Ullsperger, M. (2011). Post-error adjustments. *Frontiers in Psychology*,
700 2(September), 233. <https://doi.org/10.3389/fpsyg.2011.00233>
- 701 De Gardelle, V., & Summerfield, C. (2011). Robust averaging during perceptual judgment.
702 *Proceedings of the National Academy of Sciences of the United States of America*,
703 108(32), 13341–13346. <https://doi.org/10.1073/pnas.1104517108>
- 704 Desender, K., Boldt, A., Verguts, T., & Donner, T. H. (2019). Confidence predicts speed-
705 accuracy tradeoff for subsequent decisions. *ELife*, 8. <https://doi.org/10.7554/eLife.43499>
- 706 Desender, K., Van Opstal, F., & Van Den Bussche, E. (2017). Subjective experience of
707 difficulty depends on multiple cues. *Scientific Reports*, 7(March), 1–14.
708 <https://doi.org/10.1038/srep44222>
- 709 Drescher, L. H., Van den Bussche, E., & Desender, K. (2018). Absence without leave or
710 leave without absence: Examining the interrelations among mind wandering,
711 metacognition and cognitive control. *PLoS ONE*, 13(2), 1–18.
712 <https://doi.org/10.1371/journal.pone.0191639>
- 713 Dunlosky, J., & Connor, L. T. (1997). Age differences in the allocation of study time account
714 for age differences in memory performance. *Memory and Cognition*, 25(5), 691–700.
715 <https://doi.org/10.3758/BF03211311>
- 716 Dunn, T. L., & Risko, E. F. (2016). Toward a Metacognitive Account of Cognitive
717 Offloading. *Cognitive Science*, 40(5), 1080–1127. <https://doi.org/10.1111/cogs.12273>

- 718 Efklides, A. (2008). Metacognition. *European Psychologist*, *13*(4), 277–287.
719 <https://doi.org/10.1027/1016-9040.13.4.277>
- 720 Fernandez-Duque, D., Baird, J. A., & Posner, M. I. (2000). Executive attention and
721 metacognitive regulation. *Consciousness and Cognition*, *9*, 288–307.
722 <https://doi.org/10.1006/ccog.2000.0447>
- 723 Fisher, P. L., & Wells, A. (2008). Metacognitive therapy for obsessive-compulsive disorder:
724 A case series. *Journal of Behavior Therapy and Experimental Psychiatry*, *39*(2), 117–
725 132. <https://doi.org/10.1016/j.jbtep.2006.12.001>
- 726 Flavell, J. H. (1976). Metacognitive aspects of problem solving. In L. B. Resnick (Ed.), *The*
727 *nature of intelligence* (pp. 231–236). Hillsdale, NJ: Erlbaum.
- 728 Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general
729 Bayesian framework for metacognitive computation. *Psychological Review*, *124*(1), 91–
730 114. <https://doi.org/10.1037/rev0000045>
- 731 Fleming, S. M., & Dolan, R. J. (2014). The neural basis of metacognitive ability. In *The*
732 *Cognitive Neuroscience of Metacognition* (pp. 245–265). Berlin, Heidelberg: Springer.
733 https://doi.org/10.1007/978-3-642-45190-4_11
- 734 Fleming, S. M., Huijgen, J., & Dolan, R. J. (2012). Prefrontal contributions to metacognition
735 in perceptual decision making. *The Journal of Neuroscience*, *32*(18), 6117–6125.
736 <https://doi.org/10.1523/JNEUROSCI.6489-11.2012>
- 737 Fleming, J., Ownsworth, T., Doig, E., Hutton, L., Griffin, J., Kendall, M., & Shum, D. H. K.
738 (2017). The efficacy of prospective memory rehabilitation plus metacognitive skills

- 739 training for adults with traumatic brain injury: Study protocol for a randomized
740 controlled trial. *Trials*, 18(1), 1–11. doi:10.1186/s13063-016-1758-6
- 741 Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective
742 accuracy to individual differences in brain structure. *Science*, 329(5998), 1541–1543.
743 <https://doi.org/10.1126/science.1191883>
- 744 Fletcher, L., & Carruthers, P. (2012). Metacognition and reasoning. *Philosophical
745 Transactions of the Royal Society of London. Series B, Biological Sciences*, 367(1594),
746 1366–1378. <https://doi.org/10.1098/rstb.2011.0413>
- 747 Fleur, D. S., Bredeweg, B., & van den Bos, W. (2021). Metacognition: ideas and insights
748 from neuro- and educational sciences. *Npj Science of Learning*, 6(1), 13.
749 <https://doi.org/10.1038/s41539-021-00089-5>
- 750 Friston, K. J., Glaser, D. E., Henson, R. N. A., Kiebel, S., Phillips, C., & Ashburner, J.
751 (2002). Classical and Bayesian inference in neuroimaging: Applications. *NeuroImage*,
752 16(2), 484–512. <https://doi.org/10.1006/nimg.2002.1091>
- 753 Frömer, R., Nassar, M. R., Bruckner, R., Stürmer, B., Sommer, W., & Yeung, N. (2021).
754 Response-based outcome predictions and confidence regulate feedback processing and
755 learning. *ELife*, 10, 1–29. <https://doi.org/10.7554/ELIFE.62825>
- 756 Gherman, S., & Philiastides, M. (2018). Human VMPFC encodes early signatures of
757 confidence in perceptual decisions. *ELife*, 7, 1–28. <https://doi.org/10.7554/eLife.38293>
- 758 Gilbert, S. J. (2015). Strategic use of reminders: Influence of both domain-general and task-
759 specific metacognitive confidence, independent of objective memory ability.

- 760 *Consciousness and Cognition*, 33, 245–260.
761 <https://doi.org/10.1016/j.concog.2015.01.006>
- 762 Grainger, C., Williams, D. M., & Lind, S. E. (2016). Metacognitive monitoring and control
763 processes in children with autism spectrum disorder: Diminished judgement of
764 confidence accuracy. *Consciousness and Cognition*, 42, 65–74.
765 <https://doi.org/10.1016/j.concog.2016.03.003>
- 766 Guggenmos, M., Wilbertz, G., Hebart, M. N., & Sterzer, P. (2016). Mesolimbic confidence
767 signals guide perceptual learning in the absence of external feedback. *ELife*, 5, 1–19.
768 <https://doi.org/10.7554/eLife.13388>
- 769 Hauser, T. U., Allen, M., Purg, N., Moutoussis, M., Rees, G., & Dolan, R. J. (2017).
770 Noradrenaline blockade specifically enhances metacognitive performance, 1–13.
771 <https://doi.org/10.7554/eLife.24901>
- 772 Hebart, M. N., Görden, K., & Haynes, J.-D. (2015). The Decoding Toolbox (TDT): a
773 versatile software package for multivariate analyses of functional imaging data.
774 *Frontiers in Neuroinformatics*, 8(January), 1–18.
775 <https://doi.org/10.3389/fninf.2014.00088>
- 776 Hebart, M. N., Schriever, Y., Donner, T. H., & Haynes, J. D. (2014). The Relationship
777 between Perceptual Decision Variables and Confidence in the Human Brain. *Cerebral*
778 *Cortex*, 26(1), 118–130. <https://doi.org/10.1093/cercor/bhu181>
- 779 Hirose, S. (2020). Valid and powerful group statistics for decoding accuracy: Information
780 Prevalence Inference using the i -th order statistic i -test). *BioRxiv*, 578930.
781 <https://doi.org/10.1101/578930>

- 782 Hu, X., Luo, L., & Fleming, S. M. (2019). A role for metamemory in cognitive offloading.
783 *Cognition*, 193(June). <https://doi.org/10.1016/j.cognition.2019.104012>
- 784 Jentzsch, I., & Dudschig, C. (2009). Why do we slow down after an error? Mechanisms
785 underlying the effects of posterror slowing. *Quarterly Journal of Experimental*
786 *Psychology*, 62(2), 209–218. <https://doi.org/10.1080/17470210802240655>
- 787 Kantner, J., Solinger, L. A., Grybinas, D., & Dobbins, I. G. (2018). Confidence carryover
788 during interleaved memory and perception judgments. *Memory and Cognition*.
789 <https://doi.org/10.3758/s13421-018-0859-8>
- 790 Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's
791 new in psychtoolbox-3. *Perception*, 36(14), 1–16.
- 792 Kluwe, R. H. (1982). Cognitive knowledge and executive control. In D. Griffin (Ed.), *Human*
793 *mind – animal mind* (pp. 201–224). New York: Springer.
- 794 Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural
795 coding and computation. *Trends in Neurosciences*, 27(12), 712–719.
796 <https://doi.org/10.1016/j.tins.2004.10.007>
- 797 Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of
798 knowing. *Psychological Review*, 100(4), 609–639. [https://doi.org/10.1037/0033-](https://doi.org/10.1037/0033-295X.100.4.609)
799 [295X.100.4.609](https://doi.org/10.1037/0033-295X.100.4.609)
- 800 Koriat, A., Ackerman, R., Adiv, S., Lockl, K., & Schneider, W. (2014). The effects of goal-
801 driven and data-driven regulation on metacognitive monitoring during learning: a
802 developmental perspective. *Journal of Experimental Psychology: General*, 143(1), 386–
803 403. <https://doi.org/10.1037/a0031768>

- 804 Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between
805 monitoring and control in metacognition: Lessons for the cause-and-effect relation
806 between subjective experience and behavior. *Journal of Experimental Psychology:*
807 *General*, 135(1), 36–69. <https://doi.org/10.1037/0096-3445.135.1.36>
- 808 Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain
809 mapping. *Proceedings of the National Academy of Sciences of the United States of*
810 *America*, 103(10), 3863–3868. <https://doi.org/10.1073/pnas.0600244103>
- 811 Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., & Baker, C. I. (2009). Circular analysis
812 in systems neuroscience: The dangers of double dipping. *Nature Neuroscience*, 12(5),
813 535–540. <https://doi.org/10.1038/nn.2303>
- 814 Lak, A., Okun, M., Moss, M. M., Kepecs, A., Harris, K. D., Carandini, M., ... Carandini, M.
815 (2020). Dopaminergic and Prefrontal Basis of Learning from Sensory Confidence and
816 Reward Value Article Dopaminergic and Prefrontal Basis of Learning from Sensory
817 Confidence and Reward Value. *Neuron*, 1–12.
818 <https://doi.org/10.1016/j.neuron.2019.11.018>
- 819 MacDonald, A. W., Cohen, J. D., Stenger, V. A., & Carter, C. S. (2000). Dissociating the
820 Role of the Dorsolateral Prefrontal and Anterior Cingulate Cortex in Cognitive Control.
821 *Science*, 288(5472), 1835–1838. <https://doi.org/10.1126/science.288.5472.1835>
- 822 Masset, P., Ott, T., Lak, A., Hirokawa, J., & Kepecs, A. (2020). Behavior- and Modality-
823 General Representation of Confidence in Orbitofrontal Cortex. *Cell*, 1–15.
824 <https://doi.org/10.1016/j.cell.2020.05.022>

- 825 McCurdy, L. Y., Maniscalco, B., Metcalfe, J., Liu, K. Y., de Lange, F. P., & Lau, H. (2013).
826 Anatomical Coupling between Distinct Metacognitive Systems for Memory and Visual
827 Perception. *The Journal of Neuroscience : The Official Journal of the Society for*
828 *Neuroscience*, 33(5), 1897–1906. <https://doi.org/10.1523/JNEUROSCI.1890-12.2013>
- 829 Mei, N., Rankine, S., Olafsson, E., & Soto, D. (2020). Similar history biases for distinct
830 prospective decisions of self-performance. *Scientific Reports*, 10(1), 1–13.
831 <https://doi.org/10.1038/s41598-020-62719-z>
- 832 Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to
833 study choice. *Psychonomic Bulletin & Review*, 15(1), 174–179.
834 <https://doi.org/10.3758/PBR.15.1.174>
- 835 Miyamoto, K., Setsuie, R., Osada, T., & Miyashita, Y. (2018). Reversible Silencing of the
836 Frontopolar Cortex Selectively Impairs Metacognitive Judgment on Non-experience in
837 Primates. *Neuron*, 97(4), 980-989.e6. <https://doi.org/10.1016/j.neuron.2017.12.040>
- 838 Morales, J., Lau, H., & Fleming, S. M. (2018). Domain-General and Domain-Specific
839 Patterns of Activity Supporting Metacognition in Human Prefrontal Cortex. *The Journal*
840 *of Neuroscience*, 38(14), 2360–17. <https://doi.org/10.1523/JNEUROSCI.2360-17.2018>
- 841 Morey, R. D. (2008). Confidence Intervals from Normalized Data: A correction to Cousineau
842 (2005). *Tutorials in Quantitative Methods for Psychology*, 4(2), 61–64.
843 <https://doi.org/10.20982/tqmp.04.2.p061>
- 844 Moritz, S., & Woodward, T. S. (2007). Metacognitive training in schizophrenia: From basic
845 research to knowledge translation and intervention. *Current Opinion in Psychiatry*,
846 20(6), 619–625. <https://doi.org/10.1097/YCO.0b013e3282f0b8ed>

- 847 Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new
848 findings. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in*
849 *research and theory* (pp. 125–173). San Diego, CA: Academic Press.
- 850 Notebaert, W., Houtman, F., Opstal, F. Van, Gevers, W., Fias, W., & Verguts, T. (2009).
851 Post-error slowing: An orienting account. *Cognition*, *111*(2), 275–279.
852 <https://doi.org/10.1016/j.cognition.2009.02.002>
- 853 Odegaard, B., Grimaldi, P., Cho, S. H., Peters, M. A. K., Lau, H., & Basso, M. A. (2018).
854 Superior colliculus neuronal ensemble activity signals optimal rather than subjective
855 confidence. *Proceedings of the National Academy of Sciences*, 201711628.
856 <https://doi.org/10.1073/pnas.1711628115>
- 857 Qiu, L., Su, J., Ni, Y., Bai, Y., Zhang, X., Li, X., & Wan, X. (2018). The neural system of
858 metacognition accompanying decision-making in the prefrontal cortex. *PLoS Biology*,
859 *16*(4), e2004037. <https://doi.org/10.1371/journal.pbio.2004037>
- 860 Rabbitt, P. M. (1966). Errors and error correction in choice-response tasks. *Journal of*
861 *Experimental Psychology*, *71*(2), 264–272. <https://doi.org/10.1037/h0022853>
- 862 Redshaw, J., Vandersee, J., Bulley, A., & Gilbert, S. J. (2018). Development of children’s use
863 of external reminders for hard-to-remember intentions. *Child Development*, *89*(6),
864 2099–2108. <https://doi.org/10.1111/cdev.13040>
- 865 Ridderinkhof, K. R., Ullsperger, M., Crone, E. A., & Nieuwenhuis, S. (2004). The role of the
866 medial frontal cortex in cognitive control. *Science*, *306*(5695), 443–447.
867 <https://doi.org/10.1126/science.1100301>

- 868 Risko, E. F., & Gilbert, S. J. (2016). Cognitive Offloading. *Trends in Cognitive Sciences*,
869 20(9), 676–688. <https://doi.org/10.1016/j.tics.2016.07.002>
- 870 Rouault, M., McWilliams, A., Allen, M. G., & Fleming, S. M. (2018). Human Metacognition
871 Across Domains: Insights from Individual Differences and Neuroimaging. *Personality*
872 *Neuroscience*, 1, 1–13. <https://doi.org/10.1017/pen.2018.16>
- 873 Rounis, E., Maniscalco, B., Rothwell, J., Passingham, R., & Lau, H. (2010). Theta-burst
874 transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual
875 awareness. *Cognitive Neuroscience*, 1(3), 165–175.
876 <https://doi.org/10.1080/17588921003632529>
- 877 Schulz, L., Fleming, S. M., & Dayan, P. (2021). Metacognitive computations for information
878 search: Confidence in control. *BioRxiv*. <https://doi.org/10.1101/2021.03.01.433342>
- 879 Seow, T. X. F., Rouault, M., Gillan, C. M., & Fleming, S. M. (2021). How local and global
880 metacognition shape mental health. *Biological Psychiatry*.
881 <https://doi.org/10.1016/j.biopsych.2021.05.013>
- 882 Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Supra-personal
883 cognitive control and metacognition. *Trends in Cognitive Sciences*, 18(4), 186–193.
884 <https://doi.org/10.1016/j.tics.2014.01.006>
- 885 Shekhar, M., & Rahnev, D. (2018). Distinguishing the Roles of Dorsolateral and Anterior
886 PFC in Visual Metacognition. *The Journal of Neuroscience*, 38(22), 5078–5087.
887 <https://doi.org/10.1523/JNEUROSCI.3484-17.2018>

- 888 Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: an
889 integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217–240.
890 <https://doi.org/10.1016/j.neuron.2013.07.007>
- 891 Shimamura, A. P. (2000). The role of the prefrontal cortex in dynamic filtering.
892 *Psychobiology*, *28*(2), 207–218. <https://doi.org/10.3758/BF03331979>
- 893 Son, L. K., & Schwartz, B. L. (2009). The relation between metacognitive monitoring and
894 control. *Applied Metacognition*, 15–38. <https://doi.org/10.1017/cbo9780511489976.003>
- 895 Vaccaro, A. G., & Fleming, S. M. (2018). Thinking about thinking: A coordinate-based meta-
896 analysis of neuroimaging studies of metacognitive judgements. *Brain and Neuroscience*
897 *Advances*, *2*, 239821281881059. <https://doi.org/10.1177/2398212818810591>
- 898 Vaghi, M. M., Luyckx, F., Sule, A., Fineberg, N. A., Robbins, T. W., & De Martino, B.
899 (2017). Compulsivity Reveals a Novel Dissociation between Action and Confidence.
900 *Neuron*, 1–7. <https://doi.org/10.1016/j.neuron.2017.09.006>
- 901 Wokke, M. E., Achoui, D., & Cleeremans, A. (2020). Action information contributes to
902 metacognitive decision-making. *Scientific Reports*, *10*(1), 1–15.
903 <https://doi.org/10.1038/s41598-020-60382-y>
- 904 Ye, Q., Zou, F., Lau, H., Hu, Y., & Kwok, S. C. (2018). Causal Evidence for Mnemonic
905 Metacognition in Human Precuneus. *The Journal of Neuroscience*, *38*(28), 6379–
906 6387. <https://doi.org/10.1523/JNEUROSCI.0660-18.2018>
- 907 Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The neural basis of error detection:
908 Conflict monitoring and the error-related negativity. *Psychological Review*, *111*(4),
909 931–959. <https://doi.org/10.1037/0033-295X.111.4.939>

- 910 Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making:
911 Confidence and error monitoring. *Philosophical Transactions of the Royal Society B:*
912 *Biological Sciences*, 367(1594), 1310–1321. <https://doi.org/10.1098/rstb.2011.0416>
- 913 Yeung, N., & Summerfield, C. (2014). Shared mechanisms for confidence judgements and
914 error detection in human decision making. In S. M. Fleming & C. D. Frith (Eds.), *The*
915 *cognitive neuroscience of metacognition* (pp. 147–167). Berlin, Heidelberg: Springer.
916 https://doi.org/10.1007/978-3-642-45190-4_7
- 917 Yokoyama, O., Miura, N., Watanabe, J., Takemoto, A., Uchida, S., Sugiura, M., ...
918 Nakamura, K. (2010). Right frontopolar cortex activity correlates with reliability of
919 retrospective rating of confidence in short-term recognition memory performance.
920 *Neuroscience Research*, 68(3), 199–206.
921 <https://doi.org/10.1016/j.neures.2010.07.2041>
- 922
- 923
- 924
- 925
- 926
- 927
- 928
- 929

930 **Author contributions**

931 AB and SJG designed the research, collected and analyzed the data, and wrote the
932 manuscript. AB prepared the figures. Both authors read and approved the final manuscript.

933

934 **Figure Legends**

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

956

957 Figure 2: A) Target detection accuracy across the range of twelve target colors, placed
958 equidistant in RGB space. Only trials without a reminder were included and the data were

959 averaged across conditions (metacognitive monitoring and metacognitive control). The thick
960 black line indicates the sample average, whereas thinner lines represent individual
961 participants. B) Placement of the cursor in the Baseline condition relative to the marked
962 position on the scale (shown in red). Each black line indicates the data from a single
963 participant. C) Target detection accuracy as a function of miniblock condition (Metacognitive
964 Monitoring and Metacognitive Control) and offloading condition (Own Memory and
965 Reminder) as a raincloud plot (Allen et al., 2019). The distributions depict the densities of the
966 participant-wise accuracy averages for each condition. The boxplots show the median and
967 interquartile range of the data and the lines represent the individual observations.

968

969 Figure 3: A) Participants' ratings of Metacognitive Monitoring (low to high confidence) and
970 Metacognitive Control (high to low need for a reminder; to match orientation of the
971 confidence scale) showed a similar pattern across the twelve different colors placed
972 equidistant in RGB space. B and C) Histogram of the B) ratings and C) rating RTs pooled
973 across all participants as a function of rating type. D) Metacognitive ratings as a function of
974 objective accuracy and condition shown as participant-wise averages (left panels) and
975 correlations (right panels). MetaM = metacognitive monitoring; MetaC = metacognitive
976 control.

977

978 Figure 4: A) and B) show the regions of significant signal change in the contrast of (MetaM +
979 MetaC) > Baseline. A) Significant results were plotted on sagittal ($x = -3$), coronal ($y = 27$)
980 and axial ($z = 23$) views of the skull-stripped, mean, normalized structural image. B) Three-
981 dimensional renderings of results on right hemisphere, left hemisphere, and superior views.
982 C) Percent signal change for both metacognitive rating conditions in comparison with the
983 Baseline condition, in regions of interest (ROIs) defined by the contrasts shown in Table 2.

984 MetaM = metacognitive monitoring; MetaC = metacognitive control; r = right; l = left; MFG
985 = middle frontal gyrus; IFG = inferior frontal gyrus; SFG = superior frontal gyrus; SMA =
986 supplementary motor area. Error bars indicate +/- within-subject confidence intervals (95%)
987 according to Morey (2008).

988

989 Figure 5: A) and B) show the regions of significant signal change in the contrast of Baseline
990 > (MetaM + MetaC). A) Significant results were plotted on sagittal ($x = 0$), coronal ($y = -20$)
991 and axial ($z = 0$) views of the skull-stripped, mean, normalized structural image. B) Three-
992 dimensional renderings of results on right hemisphere, left hemisphere, and superior views.
993 C) Percent signal change for both metacognitive rating conditions in comparison with the
994 Baseline condition, in regions of interest (ROIs) defined by the task-negative contrasts shown
995 in Table 3. MetaM = metacognitive monitoring; MetaC = metacognitive control r = right; l =
996 left; SMA = supplementary motor area; MTG = middle temporal gyrus; ITG = inferior
997 temporal gyrus; OcG = occipital gyri. Error bars indicate +/- within-subject confidence
998 intervals (95%) according to Morey (2008).

999

1000 Figure 6: A) and B) show the above-chance decoding accuracy maps for the condition-
1001 specific classification analyses (blue: train on MetaM, test on MetaM; red: train on MetaC,
1002 test on MetaC; yellow: train on MetaC, test on MetaM). A) Significant results were plotted
1003 on sagittal ($x = -5$), coronal ($y = 7$) and axial ($z = 43$) views of the skull-stripped, mean,
1004 normalized structural image. B) Three-dimensional renderings of results on right hemisphere,
1005 left hemisphere, and superior views. C) and D) show the above-chance decoding accuracy
1006 maps when all four classification analyses were averaged (condition-blind decoding). C)
1007 Significant results were plotted on sagittal ($x = -18$), coronal ($y = 38$) and axial ($z = 3$) views
1008 of the skull-stripped, mean, normalized structural image. D) Three-dimensional renderings of

1009 results on right hemisphere, left hemisphere, and superior views. E) Above-chance
1010 classification accuracy for all four classification analyses (trained and/or tested on MetaM
1011 and MetaC, respectively) in regions of interest (ROIs) defined by a condition-blind selection
1012 contrast that averaged across all four analyses, listed in Table 5. MetaM = metacognitive
1013 monitoring; MetaC = metacognitive control, r = right; l = left; SFG = superior frontal gyrus;
1014 MFG (medial frontal gyrus). Error bars indicate +/- within-subject confidence intervals
1015 (95%) according to Morey (2008).

1016 **Tables**

1017 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

1018

1019 Table 2: Regions of increased signal in the MetaM and MetaC conditions, relative to the

1020 Baseline condition. MetaM = metacognitive monitoring; MetaC = metacognitive control; l =

1021 left; r = right.

Contrast	Label	Laterality	Peak voxel MNI co-ordinates	k_E	p_{FWE} cluster-corrected	Z_{max} at peak level
(MetaM + MetaC) > Baseline	Occipital and parietal cortex (calcarine cortex; cuneus; precuneus;	right and left	30, -55, 5	1338	< 0.001	5.14

	lateral ventricles; all regions both l and r)					
	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
	Pre- and postcentral gyri	right	18, -28, 65	75	0.046	4.08
	Angular gyrus	right	57, -58, 44	87	0.028	4.05

1022

1023 Table 3: Regions of decreased signal in the MetaM and MetaC conditions, relative to the
 1024 Baseline condition. MetaM = metacognitive monitoring; MetaC = metacognitive control; l =
 1025 left; r = right.

Contrast	Label	Laterality	Peak voxel MNI co-	k_E	p_{FWE} cluster-corrected	Z_{max} at peak level
-----------------	--------------	-------------------	---------------------------	----------------------	--	--------------------------------------

			ordinates			
Baseline > (MetaM + MetaC)	Cingulate and paracingulate cortices; SMA (supplementary motor area; all regions both r and l)	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
	MTG and ITG (middle and inferior temporal gyri); OcG (occipital gyri)	left	-45, -61, 8	534	< 0.001	4.93
	MTG and ITG (middle and inferior temporal gyri); OcG (occipital gyri)	right	57, -55, -4	508	< 0.001	4.92
	Anterior	right	3, 32, -4	560	< 0.001	4.61

	cingulate gyrus					
--	-----------------	--	--	--	--	--

1026

1027 Table 4: Clusters of above-chance classification accuracy in the four classification analyses.

1028 MetaM = metacognitive monitoring; MetaC = metacognitive control; l = left; r = right.

MVPA	Label	Laterality	Peak voxel MNI co-ordinates	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 l and r)	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 l and r)	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
MetaC (low vs. high need for a reminder)	Occipital pole	left	-9, -100, 14	93	< 0.001	4.09
	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
MetaC → MetaM	Superior and middle frontal gyri	left	-21, 11, 62	52	< 0.001	4.14

MetaM →	/	/	/	/	/	/
MetaC						

1029

1030 Table 5: Clusters of above-chance classification accuracy in the condition-blind classification

1031 analyses.

Label	Laterality	Peak voxel MNI co- ordinates	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;	left	-15, -70, 44	91	0.002	3.65

1032

precuneus)					
------------	--	--	--	--	--











