scientific reports

Check for updates

OPEN

Elucidating the underlying components of metacognitive systematic bias in the human dorsolateral prefrontal cortex and inferior parietal cortex

Peiyao Cong, Yiting Long, Xiaojing Zhang, Yanlin Guo & Yingjie Jiang✉

Metacognitive systematic bias impairs human learning efficiency, which is characterized by the inconsistency between predicted and actual memory performance. However, the underlying mechanism of metacognitive systematic bias remains unclear in existing studies. In this study, we utilized judgments of learning task in human participants to compare the neural mechanism difference in metacognitive systematic bias. Participants encoded words in fMRI sessions that would be tested later. Immediately after encoding each item, participants predicted how likely they would remember it. Multivariate analyses on fMRI data demonstrated that working memory and uncertainty decisions are represented in patterns of neural activity in metacognitive systematic bias. The available information participants used led to overestimated bias and underestimated bias. Effective connectivity analyses further indicate that information about the metacognitive systematic bias is represented in the dorsolateral prefrontal cortex and inferior parietal cortex. Different neural patterns were found underlying overestimated bias and underestimated bias. Specifically, connectivity regions with the dorsolateral prefrontal cortex, anterior cingulate cortex, and supramarginal gyrus form overestimated bias, while less regional connectivity forms underestimated bias. These findings provide a mechanistic account for the construction of metacognitive systematic bias.

Keywords Metacognition; Systematic bias; fMRI; MVPA

A fundamental issue in human memory research is the relationship between objective (memory) and subjective (metacognitive monitoring) dimensions of memory. The relationship between objective and subjective memory processes can be studied during learning (memory encoding at study) or retrieval (recall or recognition at test). Subjective or metacognitive monitoring during learning is of particular interest because these processes can enhance learning effectiveness by guiding the allocation of resources at a time when information remains available for learning. As a typical example of metacognitive monitoring, judgments of Learning (JOLs) are individuals' assessments of the likelihood that currently learned items will be successfully retrieved on subsequent tests, usually occurring after learning and before testing[1,2]. A common occurrence in judgments of Learning (JOLs) is that individuals tend to overestimate their ability to recall information learned during the learning phase, yet fail to recall it during a subsequent memory test. This overestimation bias is a significant issue. Conversely, there are instances where individuals underestimate their recall ability, yet perform successfully during the memory test. This is known as the underestimate bias. Numerous behavioral studies have observed this inconsistency between predicted and actual memory performance[3–7]. However, the underlying neural mechanisms remain unclear. This study aims to address this gap by exploring the neural basis of systematic biases in metacognitive monitoring. It emphasizes the inconsistency between memory predictions and actual performance, encompassing both overestimating and underestimating biases. The dual-memory monitoring hypothesis posits that making judgments of learning requires information from both working and episodic memories[8]. Alternatively, the memory strength hypothesis suggests that judgments of learning are based on the strength of working memory[9,10]. The monitoring dual-memories hypothesis and memory strength hypothesis

School of Psychology, Northeast Normal University, 5268 Renmin Street, Changchun, Jilin 130024, China. ✉email: jiangyj993@nenu.edu.cn

offer partial explanations for systematic biases in metacognitive monitoring. The information stored in working memory is unstable, leading participants to make biased memory predictions based on inaccurate information, thus creating a metacognitive systematic bias. Understanding this metacognitive systematic bias is crucial, as the neural representations of working memory can provide neural evidence for previous theoretical hypotheses.

Neural mechanism research has focused on the underlying neural substrates of metacognitive monitoring, exploring associated brain regions such as the ventromedial prefrontal cortex (vmPFC), dorsomedial prefrontal cortex (dmPFC), and anterior cingulate cortex (ACC)[11–13]. Although researchers have identified the localization of metacognitive monitoring, there is a lack of neural evidence for metacognitive systematic bias. Previous fMRI studies have found that each participant exhibits both underestimated bias and overestimated bias through the relationship between estimate memory performance (JOLs) and actual memory performance. Specifically, high JOLs magnitudes that fail to predict recall represent overestimated biases, while low JOLs magnitudes that accurately predict recall represent underestimated biases[14]. This classification helps to investigate the neural substrates underlying predicted memory outcomes (JOLs) compared to actual memory outcomes (memory itself). This study utilized the fMRI technique, combining univariate analysis methods with multivariate pattern analysis, to observe the neural patterns of systematic bias and clarify the mechanisms of overestimated bias or underestimated bias.

From the perspective of episodic memory, studies have primarily focused on behavioral mechanisms of systematic bias, specifically, overestimating or underestimating episodic memory outcomes in tasks such as color generation, emotion experience, and future events prediction[15–18]. Metacognitive monitoring is a crucial factor influencing episodic memory, playing a pivotal role in understanding the neural mechanisms underlying systematic bias. This knowledge is crucial in enhancing our understanding of efficient learning.

In this study, we employed a classic paradigm of metacognitive monitoring, in which participants encoded word pairs and provided judgments of learning (JOLs). The JOLs paradigm we used featured cue words on the left and a question mark "?" on the right. This setup is believed to prompt retrieval attempts among individuals, as observed in previous behavioral studies on immediate judgments of learning[6,19,20]. Since immediate judgments of learning occur immediately after encoding, essentially involving immediate retrieval attempts, the relevant information remains in working memory. Previous theories highlight that immediate judgments of learning may incorporate working memory information[9,21]. Therefore, it can be inferred that individuals utilize working memory information to formulate their JOLs. Furthermore, the immediate judgments of learning paradigm is similar to working memory paradigms, as both involve judgments made shortly after encoding. Although the neural mechanisms underlying immediate JOLs are not the primary focus of this study, they still need further exploration and investigation. Word pairs have been a frequent choice for metacognitive systematic bias research[3,6,22,23], often including nouns[22,24]. By selecting word pairs as learning materials, we were able to delve into the neural basis of metacognitive systematic bias. Our research objectives were threefold. First, we aimed to identify the brain regions associated with metacognitive systematic bias. To achieve this, we used univariate analysis to compare neural activation patterns between overestimated and underestimated biases during the JOLs task. Second, we sought to decode the brain regions that encode metacognitive systematic bias. To do so, we employed multivariate pattern analysis to identify brain regions that encoded information about overestimated and underestimated biases. Our third objective was to investigate the neural network of metacognitive systematic bias. The metacognitive network has been studied for over five years[25]. Despite this, there is still limited knowledge about the neural substrates of metacognitive monitoring. It is crucial to utilize effective connectivity analysis to observe neural networks that overestimate or underestimate bias. This neural evidence is significant as it provides valuable insights into the neural substrates associated with metacognitive systematic bias. Specifically, it aids in the construction of a metacognitive brain network that can further our understanding of systematic bias.

Experiment

Methods

Participants

The sample size in the current study was roughly determined by following previous study using a similar task paradigm[26]. 20 subjects participated in the experiments conducted in the current study. All participants were right-handed, had normal visual acuity or corrected visual acuity, and had no personal or family history of neurological or psychiatric disorders based on their self-report. This experiment was approved by the ethics committee of Northeast Normal University. The present study was in agreement with the Helsinki Declaration and approved by the ethics committee of the Northeast Normal University (Study No. 2022020).The participants signed an informed consent form before the experiment and were paid for completing the experiment and received a payment of 100 CNY once the experiment was completed.

Stimuli

The 126 abstract word pairs from Yu, Jiang, and Li (2020) were used, and each item is middle difficulty (0.3 to 0.7) through a memory recognition task[27]. Among them, 120 pairs of words were used for the formal experiment, and the remaining six pairs of abstract words were used for practice.

Procedure

In this study, we used an event-related design (Fig. 1). Figure 1 showed the details of the procedure. The formal scan consisted of 4 runs, with a short break given to the participants at the end of each run. The task took approximately 30 min to complete inside the scanner. In the scanner and each run, participants performed an encoding and immediate judgment of learning (JOLs) task. During the encoding and immediate JOLs

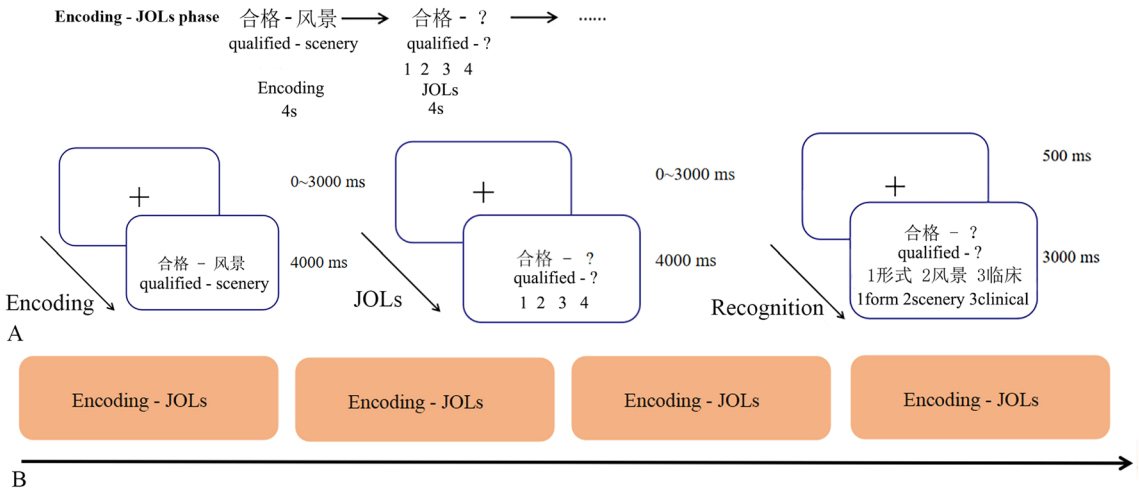

Figure 1. Experiment paradigm. (**A**) The rapid event-related design was used to fit the encoding-JOLs phase better. The major procedure in fMRI contained encoding-JOLs phase. Recognition-test phase was outside the fMRI scanner. Also, the details of each typical trial were introduced. (**B**) The arrangement of scanning runs. There were four encoding-JOLs sessions in total.

stage, participants saw random jitters on the center of the screen ranging from 0 to 4000 ms, followed by the presentation of an abstract word pair (e.g., "合格-风景", written in Latin characters "qualified—scenery") to be learned for 4000 ms (total 16 word pairs). After encoding each pair, participants saw one word from the pair (the cue) on the screen and were asked to predict how likely they would remember the unseen target in the post-scan recognition task on a four-point scale, with 1 indicating "will be absolutely forgotten" and 4 indicating "will be absolutely remembered". Participants had 4000 ms to press a button to indicate their estimated performance, and responses were collected online using an MRI-compatible button box. After the encoding-JOLs, a distraction task outside the scanner was asked to complete for 3 min. Participants also were not in the scanner during the recognition-test phase. In a recognition test trial, participants saw a previous cue word that was studied at the top of the screen, and the target word and two distractor words appeared in random locations (left, center, or right) on the bottom. Participants were asked to indicate which of the three words on the bottom had been paired initially with the cue at the top in 3000 ms. Each trial was associated with a fixed-interval fixation of 500 ms.

fMRI data acquisition

Neuroimaging data were acquired on a UIH Prisma 3.0 T MRI scanner with a 64-channel head-neck coil. The participant was placed in a supine position with a sponge pad inside the coil to hold the head in place and was asked to keep the head and body still during the scanning process. The functional image was a 32-slice axial image, measured by a T1 -weighted echo-planar images (EPI) sequence, covering the entire cerebral cortex (main technical parameters: TR = 2000 ms, TE = 30 ms, Flip angle = 80°, FOV = 230 mm × 230 mm, Matrix size = 64 × 64, slice thickness = 3.5 mm, sequential acquisition = 32 axial slices, voxel size = 3.5 × 3.5 × 4.2 mm). Each functional scanning session contained 207 time points, with a total of 4 runs. Structural images were collected using a T1-weighted 3D MPRAGE sequence (TR = 7 ms, TE = 3 ms, Flip angle = 9°, FOV = 230 mm × 230 mm, Matrix size = 384 × 384, slice thickness = 1 mm, sequential acquisition = 160 axial slices, voxel size = 0.5 × 0.5 × 0.5 mm), in order to coregister with the functional images.

fMRI data preprocessing

Imaging analysis was performed using spm12 (http://www.fil.ion.ucl.ac.uk/spm)[28]. First, all the EPI DICOM data were converted to NIFTI format. The first three images from each run were automatically discarded by the scanner to allow scanner equilibrium. Second, all volume slice scan times were corrected to the middle time slice and realigned to the first scan to correct for head motion. Third, the structural images of each subject were coregistered with the mean functional images, and then the images were normalized to the Montreal Neurological Institute template. Fourth, all voxels were resampled to 3 × 3 × 3 mm. Last, all functional volumes were smoothed by using an 8-mm FWHM isotropic Gaussian kernel.

Behavioral data analysis

Using participants' responses on the post-scan recognition test, we sorted trials based on JOLs magnitude and recognition performance. At the JOLs stage, participants were required to make immediate JOLs using a 1–4 scale. The 1 and 2 indicate that the participant will forget, while 3 and 4 indicate that the participant will remember. The four-point scale was used to fit the fMRI environment and was based on previous fMRI studies[29]. In the post-scan recognition test, a correct recognition was recorded as 1, and a failed recognition or timeout was recorded as 0. Therefore, items were given either an R (will remember) or an F (will forget) estimation in the JOLs stage and were either subsequently remembered (r) or subsequently forgotten (f) in the post-scan recognition memory test. This study aimed to investigate metacognitive systematic bias by comparing overestimated bias to underestimated

bias. The classification of metacognitive systematic bias is of great importance. We classified metacognitive systematic bias into two types[14]: (1) JOLs magnitude was high ("will remember" prediction) but was later failed to recognize in the post-scan recognition test ($JOL_{high}M_{low}$), which is overestimated bias. (2) JOLs magnitude was low ("will forget" prediction) but was later correctly recognized in the post-scan recognition test ($JOL_{low}M_{high}$), which is underestimated bias. Regarding the classification of the 4-point scale data into two categories, this decision was a deliberate choice, tailored specifically to fulfill the research objectives of elucidating the neural mechanisms that underlie metacognitive biases. By organizing the data into two representative categories, the authors intended to pinpoint and contrast the neural disparities between the two types of metacognitive biases, ultimately disclosing their underlying formation mechanisms. The behavioral data analysis has three steps: First, we calculated response time (RT) and proportion between the $JOL_{high}M_{low}$ and $JOL_{low}M_{high}$ conditions to test the feasibility of further fMRI analysis. This step was to confirm that both overestimated bias and underestimated bias were not happening by chance. Second, metacognitive sensitivity was calculated for each participant to evaluate the overall metacognitive monitoring accuracy via meta-d/d values in accordance with Maniscalco and Lau (2012). Then, metacognitive sensitivity was calculated for both $JOL_{high}M_{low}$ and $JOL_{low}M_{high}$ conditions and should be compared between the $JOL_{high}M_{low}$ and $JOL_{low}M_{high}$ conditions. This approach can provide evidence of which type of metacognitive systematic bias is more sensitive. Because metacognitive sensitivity is an index that measures the accuracy of JOLs[30]. We have known that metacognitive systematic bias has low JOL accuracy, but it remains unclear whether overestimated bias or underestimated bias has less accuracy. Third, the recognition task performance was measured to ensure the effectiveness of the materials and tasks used in the experiment.

Univariate analysis

The GLM method, as implemented in the SPM toolbox, was used to analyze the BOLD responses to metacognitive systematic bias. For all analysis, events were modeled at the time of the stimulus onset and convolved with the canonical hemodynamic response function (HRF) using a double-gamma function. These events were then superimposed for all trials to fit with the fMRI signals of each voxel. At the JOLs stage, the event was time-locked to the onset of the stimuli, with a duration that was the summation of the presentation period (4 s) and the same duration as the event. The GLM model was based on the JOLs task. The GLM model was based on JOLs task, we separated two task-related events, including JOLs magnitude was high ("will remember" prediction) but was later failed to recognize in post-scan recognition test ($JOL_{high}M_{low}$) and JOLs magnitude was low ("will forget" prediction) but was later correct recognized in post-scan recognition test ($JOL_{low}M_{high}$). Motion correction parameters were entered as covariates of no interest, along with a constant term per run. The regressors were convolved with a canonical hemodynamic response function. Low-frequency drifts were excluded with a 1/128 Hz high-pass filter. Missed trials were not modeled. We defined two contrasts: $JOL_{high}M_{low}$ vs. $JOL_{low}M_{high}$ (1 -1), $JOL_{low}M_{high}$ vs. $JOL_{high}M_{low}$ (− 1 1). Contrasts constructed at the single participant level were then input into a second-level group analysis using a random-effects model. At the group level, metacognitive systematic bias fMRI activation was first obtained by applying a parametric one-sample *t-test*, then a paired sample *t-test* was used to compare the activation between different metacognitive systematic bias ($JOL_{high}M_{low}$ versus $JOL_{low}M_{high}$, and vice versa). All reported clusters survived a threshold with $p < 0.05$ after correcting for multiple comparisons using the false discovery rate (FDR) method and consisted of ten or more significant voxels.

Regions of interest (ROI) analysis

ROIs were defined from previous literature [11,12,14]. Voxels meeting $p < 0.05$ (FDR correction) threshold requirement and lying in the proximity of previously published coordinates of dorsomedial prefrontal cortex (dmPFC) [-6,2,58], ventromedial prefrontal cortex (vmPFC) [− 32,6,54], dorsolateral prefrontal cortex (dlPFC) [-48,24,28], and anterior cingulate cortex (ACC) [0,32,2][11,12,14] were taken to be the ROIs used in this study. Beta values were extracted from subjects' contrast images for the $JOL_{high}M_{low}$ and $JOL_{low}M_{high}$ univariate analyses, respectively.

Multivariate pattern analysis

Multivariate pattern analysis (MVPA) was performed in MATLAB using the CoSMoMVPA Toolbox (https://www.cosmomvpa.org/)[31]. According to research on the use of MVPA for decoding in the same field[12], we classified runwise beta images from GLMs modeling $JOL_{high}M_{low}$ and $JOL_{low}M_{high}$ activity patterns in ROI and whole-brain searchlight analyses. ROI MVPA was performed on normalized, non smoothed images using the ROI spheres as masks. Previous work has shown that these preprocessing steps have minimal impact on linear discriminant analysis (LDA) classification accuracy while allowing meaningful comparison across subject-specific differences in anatomy, as in standard fMRI analyses[32,33]. A single accuracy value per subject, per condition, and per ROI was extracted and used for group analysis and statistical testing. Whole-brain searchlight analyses used 3 mm-radius spheres centered around a given voxel for all voxels on spatially realigned and slice-time corrected images from each subject to create whole-brain accuracy maps. The significance of the classification accuracies of all voxels was tested using a non-parametric random permutation test ($n = 5000$) and results were corrected for multiple comparisons using the false discovery rate (FDR) approach (the significance threshold was set at $p < 0.05$).

For group-level analyses, these individual searchlight maps were spatially normalized and smoothed using a Gaussian kernel (8 mm FWHM) and entered into one-sample *t*-tests against chance accuracy[34]. Whole-brain cluster inference was performed in the same manner as in univariate analysis. For metacognitive systematic bias classifications, we conducted independent leave-one-run-out cross-validations on $JOL_{high}M_{low}$ activity patterns and $JOL_{low}M_{high}$ activity patterns. Pattern vectors from three of the four runs in each condition were used to train an LDA to predict the same classes in the vectors from the left-out run. We compared the true labels of the

left-out run with the labels predicted by the model and iterated this process for the other run to calculate a mean cross-validated accuracy independently for each condition.

Effective connectivity analysis

Dynamic Causal Modeling (DCM) is an effective connectivity analysis method for making inferences about causal relationships between brain regions. In this study, DCM was performed in SPM12 to compare brain connectivity strength between $JOL_{high}M_{low}$ and $JOL_{low}M_{high}$. Specifically, the volumes of interest (VOI) were defined based on brain regions that have significant activation in the univariate analysis and multivariate pattern analysis. In other words, only VOIs were significant in univariate analysis, and multivariate pattern analysis included DCM analysis. Within each VOI, we chose the radius of 8 mm as centers of spherical VOIs based on contrasts within a GLM: $JOL_{high}M_{low}$ versus $JOL_{low}M_{high}$ and $JOL_{low}M_{high}$ versus $JOL_{high}M_{low}$. According to previous studies[35], in DCM analysis, three parameters need to be determined: matrix A (internal parameter), matrix B (modulation parameter), and matrix C (driving input parameter). Matrix A represents the intrinsic coupling among brain regions in the absence of external perturbations, and in this study, matrix A represents the whole metacognitive systematic bias. Matrix B is the change in brain region caused by the experiment, i.e., the $JOL_{high}M_{low}$ or $JOL_{low}M_{high}$ in this study. Matrix C is the perturbation of brain activity due to external input.

Our primary interest was to estimate the quantitative differences between $JOL_{high}M_{low}$ and $JOL_{low}M_{high}$ in connectivity strength. Therefore, we focused on quantitative comparisons of the DCM parameters (in particular, matrix *B*) between $JOL_{high}M_{low}$ and $JOL_{low}M_{high}$. The full model described above was first estimated at the individual level to derive DCM parameters for hypothesis testing at the group level. Then, groups of multiple subjects were averaged using PEB (Parametric Empirical Bayes) and BMR (Bayesian Model Reduction)[35]. The posterior probability (P) > 0.95 was used to indicate the significance of the model. Pairwise tests were also performed between $JOL_{high}M_{low}$ and $JOL_{low}M_{high}$ conditions, with posterior probabilities (P) > 0.95 indicating the significance of each brain region.

Results

Behavioral results

Paired sample *t-tests* revealed no significant difference in RT and proportion between $JOL_{high}M_{low}$ ($M_{RT} = 1024.83$; $M_{proportion} = 0.23$) and $JOL_{low}M_{high}$ ($M_{RT} = 1169.64$; $M_{proportion} = 0.23$), $t_{(16)} = -1.75$, $p = 0.099$, $BF_{10} = 0.876$; $t_{(16)} = 0.039$, $p = 0.969$, $BF_{10} = 0.249$ (see Fig. 2), indicating suitable classification per systematic bias type for further fMRI analysis. Metacognitive sensitivity for each participant was measured via meta-d/d values in accordance with Signal Detection Theory[30], indicating that participants had lower metacognitive sensitivity, $M = -1.26 \pm 0.23$. Then metacognitive sensitivity of $JOL_{high}M_{low}$ and $JOL_{low}M_{high}$ were measured, and paired sample *t-tests* showed significant difference, $t_{(16)} = -4.30$, $p < 0.001$, $BF_{10} = 63.11$, means $JOL_{high}M_{low}$ have lower metacognitive sensitivity than $JOL_{low}M_{high}$. The correct recognition rate for all subjects was 56.60% ± 18%, indicating that the subjects completed the task carefully.

Univariate analysis results

Metacognitive systematic bias whole-brain responses were first analyzed using the conventional GLM method. As shown in Fig. 3A,B, $JOL_{high}M_{low}$, and $JOL_{low}M_{high}$ all activated left dlPFC and left dmPFC. Other regions activated included left supramarginal, right precuneus, right superior frontal gyrus (SFG), left middle temporal gyrus (MTG), and right superior temporal gyrus (STG) under $JOL_{high}M_{low}$ condition. We found elevated activity in ACC and left insula under $JOL_{low}M_{high}$ condition (see Fig. 3 and Table 1). Furthermore, comparing metacognitive systematic bias BOLD activation between $JOL_{high}M_{low}$ and $JOL_{low}M_{high}$ showed other regions activated included left inferior parietal lobule (IPL) and left middle cingulate cortex (MCC) in $JOL_{high}M_{low} > JOL_{low}M_{high}$ contrast, left parahippocampal in $JOL_{high}M_{low} < JOL_{low}M_{high}$ contrast (see Fig. 3C,D).

The ROI analysis results showed no significant difference between $JOL_{high}M_{low}$ and $JOL_{low}M_{high}$ in left dlPFC ($M = 0.23$, $M = 0.30$), left dmPFC ($M = 0.40$, $M = 0.35$), and left vmPFC ($M = —0.23$, $M = —0.37$). ACC were more activated in $JOL_{low}M_{high}$ than $JOL_{high}M_{low}$ condition, paired sample *t-tests* : $t_{(17)} = 4.95$, $p < 0.001$.

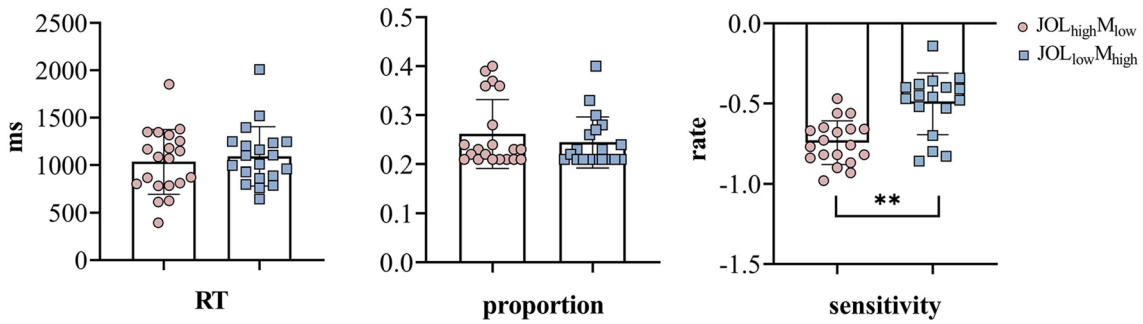

Figure 2. Behavior results in experiment 1. The left panel showed the RT results between $JOL_{high}M_{low}$ and $JOL_{low}M_{high}$. The right panel represents proportion results and metacognitive sensitivity results between $JOL_{high}M_{low}$ and $JOL_{low}M_{high}$. $^{**}p < 0.01$.

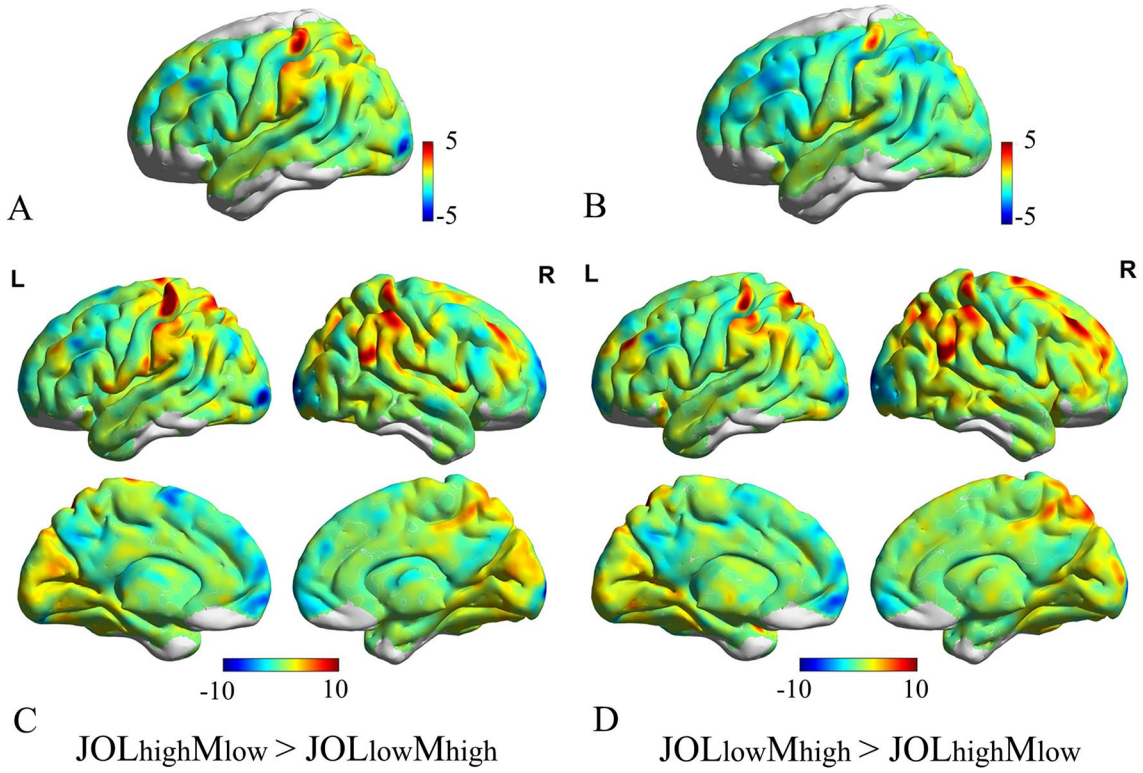

Figure 3. Univariate analysis of metacognitive systematic bias activity in experiment. (**A**) $JOL_{high}M_{low}$ activates left dlPFC, left supramarginal, right precuneus, right SFG, left MTG, right STG. (**B**) $JOL_{low}M_{high}$ activates left dlPFC, ACC, and left insula. (**C**) univariate BOLD activation in left IPL and left MCC showed a significant difference in $JOL_{high}M_{low} > JOL_{low}M_{high}$ contrast. (**D**) univariate BOLD activation in the left parahippocampal on $JOL_{high}M_{low} < JOL_{low}M_{high}$ contrast. $p < 0.05$ FDR correction.

Multivariate pattern analysis (MVPA) results

A series of MVPAs were performed to obtain activity patterns of metacognitive systematic bias when remembering abstract word pairs. If systematic bias is shared across $JOL_{high}M_{low}$ and $JOL_{low}M_{high}$, then common regions would be found in these two kinds of metacognitive systematic bias.

ROI MVPA analysis results

We performed an LDA decoding analysis using as input vectors the runwise beta images pertaining to $JOL_{high}M_{low}$ and $JOL_{low}M_{high}$ trials obtained from a GLM (12 input vectors in total). For $JOL_{high}M_{low}/JOL_{low}M_{high}$ classification, we used standard leave-one-out independent cross-validations for each condition ($JOL_{high}M_{low}/JOL_{low}M_{high}$), and we performed one sample *t-test* for each ROI and each condition, then conducted paired *t-test* for $JOL_{high}M_{low}$ versus $JOL_{low}M_{high}$ to obtain which region decoding metacognitive systematic bias information.

Mean accuracy in classifying $JOL_{high}M_{low}$ and $JOL_{low}M_{high}$ was significantly above chance level in all ROIs (one-sample *t-tests* Bonferroni corrected for multiple comparisons $\alpha = 0.05/4 = 0.0125$), shown in Fig. 4. In details, the mean accuracy of $JOL_{high}M_{low}$ in each ROI: left dlPFC, $t(16) = 21.68$, $p < 0.001$; left dmPFC, $t(16) = 20.37$, $p < 0.001$; left vmPFC, $t(16) = 4.06$, $p < 0.001$; ACC, $t(16) = 20.66$, $p < 0.001$; and $JOL_{low}M_{high}$ in each ROI: left dlPFC, $t(16) = 9.76$, $p < 0.001$; left dmPFC, $t(16) = 7.29$, $p < 0.001$; left vmPFC, $t(16) = 5.12$, $p < 0.001$; ACC, $t(16) = 15.09$, $p < 0.001$. In particular, paired *t-test* used to analyze the common regions in ROI analysis showed $JOL_{high}M_{low}$ classification accuracy was significantly different from $JOL_{low}M_{high}$ in left dlPFC ($t(16) = 21.68$, $p < 0.001$), left dmPFC ($t(16) = 14.46$, $p < 0.001$), left vmPFC ($t(16) = 5.12$, $p < 0.001$), ACC ($t(16) = 15.09$, $p < 0.001$) (see Fig. 4C). Consistent with our hypothesis, $JOL_{high}M_{low}$ and $JOL_{low}M_{high}$ representations could be decoded in parts of the PFC and temporal cortex.

Searchlight analysis results

We ran a similar decoding analysis using an exploratory whole-brain searchlight, obtaining a classification accuracy value per voxel when classifying $JOL_{high}M_{low}$ and $JOL_{low}M_{high}$. As shown in Fig. 4E,F, Consistent with our ROI results, we observed significant accuracy classification under $JOL_{high}M_{low}$ condition (Fig. 4E) in left dlPFC ($t_{(16)} = 15.97$, $p < 0.001$), left dmPFC ($t_{(16)} = 35.74$, $p < 0.001$), left vmPFC, ($t_{(17)} = 14.53$, $p < 0.001$), ACC ($t_{(16)} = 20.06$, $p < 0.001$), and significant accuracy classification under $JOL_{high}M_{low}$ condition in left dlPFC ($t_{(16)} = 25.68$, $p < 0.001$), left dmPFC ($t_{(16)} = 14.46$, $p < 0.001$), left vmPFC, ($t_{(17)} = 5.12$, $p < 0.001$), ACC ($t_{(16)} = 15.09$, $p < 0.001$) (one-sample *t-test* Bonferroni corrected for multiple comparisons $\alpha = 0.05/4 = 0.0125$). Searchlight analysis found other regions decoded $JOL_{high}M_{low}$ information (Fig. 4F), specifically, left supramarginal gyrus ($t_{(16)} = 16.33$, $p < 0.001$), right precuneus ($t_{(16)} = 20.26$, $p < 0.001$), and other regions decoded $JOL_{low}M_{high}$ information: left insula

contrast	Anatomical Region	MNI coordinates (*x*, *y*, *z*)	*Z* score	*p* value	Hemisphere
$JOL_{high}M_{low} > JOL_{low}M_{high}$	Inferior parietal lobule	− 52,− 42,56	4.40	$p < 0.001$ $p = 0.00001$	Left
	Middle cingulate cortex	− 4,− 26,36	3.71	$p < 0.001$ $p = 0.0002$	Left
$JOL_{high}M_{low} < JOL_{low}M_{high}$	Parahippocampal	− 18,− 18,− 24	4.18	$p < 0.001$ $p = 0.00003$	Left
$JOL_{high}M_{low}$	Middle cingulate cortex	− 4,− 26,36	3.35	$p < 0.001$ $p = 0.0008$	Left
	Dorsolateral prefrontal cortex	− 48,24,28	6.26	$p < 0.001$ $p = 0.000000004$	Left
	Dorsomedial prefrontal cortex	− 6,2,58	5.15	$p < 0.001$ $p = 0.0000002$	Left
	Supramarginal	− 58,− 24,28	7.09	$p < 0.001$ $p = 0.00000000001$	Left
	Precuneus	10,− 52,28	5.11	$p < 0.001$ $p = 0.0000003$	Right
	Superior frontal gyrus	8,68,14	5.2	$p < 0.001$ $p = 0.0000002$	Right
	Middle temporal gyrus	− 60,− 24,− 14	6.72	$p < 0.001$ $p = 0.00000000001$	Left
	Superior temporal gyrus	48,− 20,6	6.5	$p < 0.001$ $p = 0.00000000001$	Right
$JOL_{low}M_{high}$	Dorsolateral prefrontal cortex	− 48,24,28	5.01	$p < 0.001$ $p = 0.0000005$	Left
	Dorsomedial prefrontal cortex	− 6,2,58	4.03	$p < 0.001$ $p = 0.00005$	Left
	Anterior cingulate cortex	0,32,2	5.55	$p < 0.001$ $p = 0.0000003$	Left/Right
	Insula	− 38,− 15,7	3.69	$p < 0.001$ $p = 0.0002$	Left

Table 1. MNI coordinates and corresponding Z scores for brain areas activated by $JOL_{high}M_{low}$ and $JOL_{high}M_{low}$ conditions.

($t_{(16)} = 15.14$, $p = 0.000$), left IFG ($t_{(16)} = 10.13$, $p < 0.001$), right precuneus ($t_{(16)} = 11.26$, $p < 0.001$). Furthermore, a paired *t*-test was used to analyze the common regions in searchlight analysis and showed higher decoding accuracy for $JOL_{high}M_{low}$ than $JOL_{low}M_{high}$ in the right precuneus ($t_{(16)} = 2.73$, $p = 0.016$). These results revealed that the different part of the brain region represents information about specific metacognitive systematic bias, and common regions of PFC shared information across $JOL_{high}M_{low}$ and $JOL_{low}M_{high}$.

Effective connectivity results

Figure 5A,B shows the PEB analysis results for the modulatory effects on the effective connectivity between the modeled nodes. Connection strengths of the parameters whose posterior probability was higher than 0.95 ($P > 0.95$) are reported. The results under $JOL_{high}M_{low} > JOL_{low}M_{high}$ condition found a significant single connection from left dlPFC to right precuneus, and bidirectional connections between left dmPFC and ACC, right precuneus and ACC, left dmPFC and left supramarginal gyrus, left insula and left supramarginal gyrus.

The results under the $JOL_{low}M_{high} > JOL_{high}M_{low}$ condition showed a significant single connection from left dlPFC to left dmPFC, left insula to ACC, and bidirectional connections between left dmPFC and left supramarginal gyrus (Fig. 5B).

Discussion

A critical question in metacognitive monitoring is why individuals are sometimes inclined to overestimate or underestimate their memory performances. The neural mechanism of metacognitive systematic bias for overestimate prediction versus underestimate prediction was examined in this study using fMRI, machine learning decoding, and effective connectivity. In particular, we direct our attention on whether metacognitive brain regions and working memory regions engage in the formation of systematic bias when making JOLs. We found dissociated neural mechanisms that supported overestimated bias and underestimated bias, and the results should deepen our understanding of the cognitive and neural mechanisms of metacognitive systematic bias and thus help to answer the question of how this bias occurs.

Neural correlates of metacognitive systematic *bias*

Our results could help clarify the role of dlPFC in JOLs and the working memory process. As a typical metacognitive monitoring region, the activation of dlPFC was found in previous studies [13,14,36,37]. One possible explanation posits that increased dlPFC activity reflects partial retrieval of the target word in working memory[14], but this hypothesis is contradicted by the fact that dlPFC is more dorsal to the regions involved in semantic elaboration[29]. The debate on dlPFC was partly resolved through a TMS study. Rounis et al. (2010) found causal

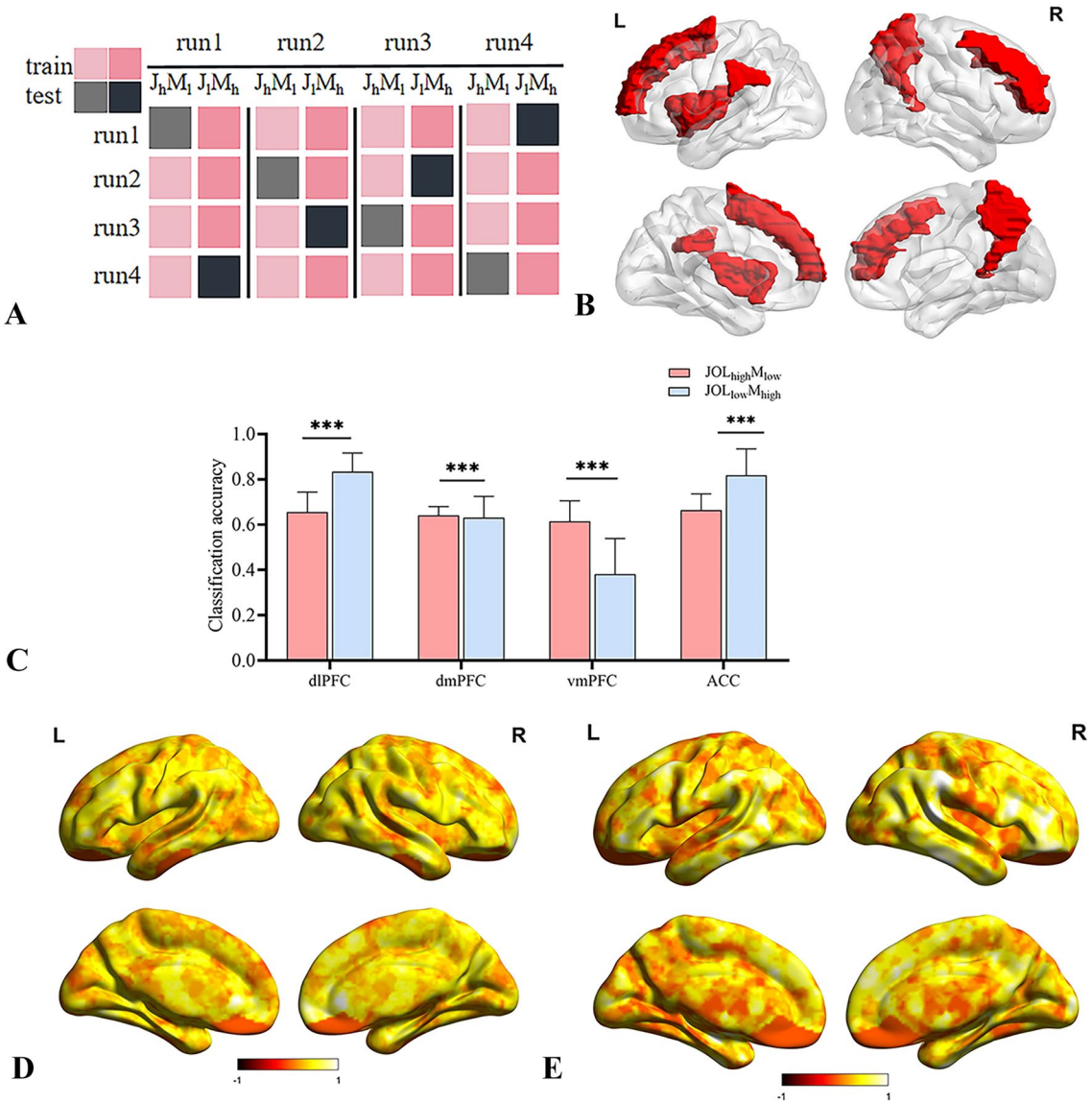

Figure 4. MVPA results. (**A**) Pattern vectors of two classes (e.g., $JOL_{high}M_{low}$ and $JOL_{low}M_{high}$) were used to train a decoder in a leave-one-run-out design that was then tested in the left-out pair. The process was iterated four times to test pairs from every run. (**B**) Mask used in ROI MVPA analysis (**C**) ROI results for $JOL_{high}M_{low}$ versus $JOL_{low}M_{high}$ classification accuracy in experiment. (**D**) Searchlight analysis results for $JOL_{high}M_{low}$ classification accuracy in experiment. (**E**) Searchlight analysis for $JOL_{low}M_{high}$ classification accuracy in experiment. ***$p < 0.001$. All clusters in D and E are significant at a cluster-based permutation test ($p < 0.05$), corrected for multiple comparisons at $p_{FDR} < 0.05$.

evidence that dlPFC TMS decreases metacognitive accuracy[38]. Using the JOLs paradigm and MVPA analysis, this study found strong evidence that dlPFC represents metacognitive monitoring. Specifically, univariate fMRI analysis showed that the JOL stage evoked metacognitive monitoring-related BOLD activity in dlPFC, and MVPA revealed that the decoding accuracy in dlPFC was significantly above the chance level in the experiment. It is suggested that dlPFC, as a metacognitive monitoring region, plays a fundamental role in the formation of metacognitive systematic bias.

Another region was found in ACC, which is known for performance monitoring[11], integration of detected conflicts[39], and attentional control mechanisms[40]. It has been shown that the cingulate cortex plays a major role in detecting discrepancies between the intended and the actual outcome of an action[41]. The significant classification accuracy of the ACC in the context of predicting memory performance (JOLs) might reflect its engagement in general performance monitoring. This result was supported by previous univariate fMRI analysis, and this study observed the ACC through machine learning decoding that supports the basic function of the ACC in the formation of metacognitive systematic bias.

As has been mentioned previously, making metacognitive monitoring predictions requires retrieval of the target word in working memory[9,10,14]. This is because at that time, the slow memory traces are weak, and participants will overestimate or underestimate their memory performance. Some regions represent the storage

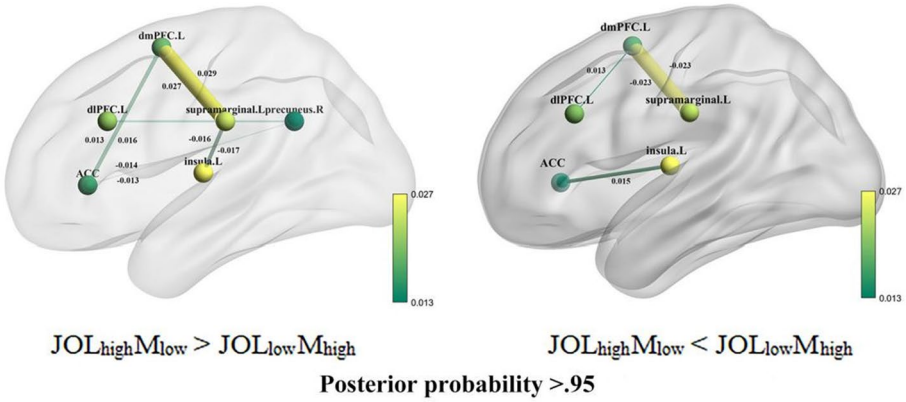

Figure 5. Effective connectivity results in experiment. (**A**) Effective connectivity results for $JOL_{high}M_{low} > JOL_{low}M_{high}$. (**B**) Effective connectivity results for $JOL_{low}M_{high} > JOL_{high}M_{low}$. Posterior probability was higher than 0.95 ($P > 0.95$).

of working memory, e.g., inferior parietal lobule (IPL), supramarginal gyrus (SMG), angular gyrus, thalamus, superior parietal lobule (SPL)[42–44]. These regions associated with working memory were found in the results of univariate analysis of the experiment. In particular, the decoding accuracy of SMG in MVPA results was significantly above the chance level, which suggested SMG as a region involved in metacognitive systematic bias.

Just as working memory retrieval is an inference in metacognitive monitoring studies[9,10], partial evidence could support the former hypothesis if working memory representations were found in metacognitive monitoring. This study detected that the working memory representation (SMG) provides critical evidence that making metacognitive monitoring predictions requires information from working memory, giving certain neural mechanism evidence to the dual-memories hypothesis and memory strength hypothesis. Furthermore, SMG not only has a single function for memory monitoring but also works in tandem with other brain regions to predict memory. The working memory trace is a possibility to produce overestimate or underestimate bias. The cognitive and neural mechanisms of overestimate bias and underestimate bias will be discussed in the brain connectivity "Results" section.

Moreover, through searchlight analysis, we discovered an interesting finding: a significantly higher decoding accuracy for $JO_{Lhigh}M_{low}$ compared to $JOL_{low}M_{high}$ within the right precuneus. This area of the brain, the precuneus, has been recognized as integral to metacognition, as supported by correlational evidence derived from functional activity analyses. For example, previous research has demonstrated a connection between metacognitive performance related to memory decisions and the precuneus[12,45,46]. Furthermore, the precuneus plays a pivotal role in retrospective confidence ratings, exhibiting greater activation when individuals express low confidence[47]. These observations suggest that the activation level of the precuneus serves as an indicator of both high and low confidence ratings. Notably, the present study focused on prospective confidence ratings and discovered that $JOL_{high}M_{low}$ decoded more information from the precuneus than $JOL_{low}M_{high}$, thereby indicating that the precuneus reflects varying levels of confidence.

Throughout various phases of memory, the precuneus exhibits distinct patterns of activity. Specifically, when individuals provide confidence ratings immediately after encoding (judgments of learning), a stronger activation pattern is observed in the precuneus for higher confidence levels, while a weaker pattern is evident for lower confidence. Conversely, when confidence ratings are made following memory testing (judgments of confidence), a greater degree of precuneus activation is associated with lower confidence levels. Not only does the current study reveal variations in precuneal activity during confidence ratings, but it also suggests that the precuneus serves as the neural foundation for metacognitive biases. Furthermore, it appears that the precuneus contributes differentially to two types of metacognitive biases. In particular, it seems to play a more significant role in overestimation biases compared to underestimation biases, resulting in stronger activation and, consequently, higher decoding accuracy. This result not only corroborates the hypothesis of the involvement of the precuneus in metacognition processes[48,49], but also strengthens the view of a domain-specificity in the assessment of metacognition[12].

The different cognitive mechanisms between overestimated *bias* and underestimated *bias*

Through the formation of overestimate prediction and underestimate prediction, we found that SMG played an important role. However, the behavioral evidence showed that overestimate bias ($JOL_{high}M_{low}$) had lower metacognitive accuracy than underestimate bias ($JOL_{low}M_{high}$), suggesting different cognitive mechanisms behind them. The effective connectivity analysis results provided a network interpretation of the metacognitive accuracy difference. It revealed that higher brain connectivity was observed between the working memory region (IPL, SMG) and uncertainty signals region (insula) in overestimated prediction. Conversely, elevated metacognitive monitoring connectivity was found in underestimate prediction. A possible explanation for the lower metacognitive accuracy in overestimate bias is that more information increases participants' confidence[50]. When making judgments of learning, individuals require more resources (e.g., working memory resources

and metacognitive monitoring resources). A series of irrelevant information can interrupt an individual's metacognitive monitoring, leading to overestimated predictions due to inflated memory performance. Conversely, when individuals have limited information, the available resources guide them to make underestimated predictions. The brain connection focuses more on metacognitive monitoring regions, providing neural network evidence for underestimated biases. Previous studies have focused on the behavioral mechanism of metacognitive systematic bias[5–7,51] and the measurement of bias using the behavioral method. However, they lack direct evidence comparing overestimated bias and underestimated bias. This study provides clear neural evidence regarding the formation of overestimated and underestimated biases and interprets the cognitive mechanism from an information availability perspective.

Dissociable neural networks supporting metacognitive systematic *bias*

When people are overconfident or underconfident in their memory predictions, dissociable neural connectivity is observed. The effective connectivity results provide evidence that the dlPFC and dmPFC play a central role in metacognitive monitoring processes, as significant connectivity was observed between the dlPFC and SMG, dmPFC, and SMG, especially for overestimate bias. The function of the dlPFC and dmPFC should be discussed in detail. Metacognitive monitoring studies have shown that the dlPFC and dmPFC are key brain regions when making metacognitive monitoring judgments[13,14,29,36,37], while executive function studies suggest that the dlPFC and dmPFC are involved in working memory processes[52,53]. Using the JOLs paradigm and MVPA analysis, we found that the dlPFC and dmPFC are correlated with metacognitive monitoring, and the SMG represents the working memory process, indicating different neural mechanisms between metacognitive monitoring and working memory. Moreover, connectivity between the PFC and parietal cortex has been implicated in metacognition and decision-making studies[13,36,54]. In studies of decision-making, the ACC, vmPFC, and insula have been found to reveal uncertainty in decision-making[55]. The connectivity between the ACC and vmPFC, as well as the ACC and insula, was found to indicate uncertain decision-making, particularly in cases of underestimated bias, across two experiments. These findings suggest that different neural substrates are involved when making overestimated or underestimated biases. It is proposed that multiple regions, including metacognitive monitoring, working memory, and uncertainty, contribute to the formation of overestimated bias, while the collaboration of uncertainty monitoring and decision-making-related brain connectivity leads to the development of underestimated bias.

Conclusion

It is concluded that the present study has found a remarkable dissociation between the neural processes that underlie overestimate bias and underestimate bias. The results of MVPA and effective connectivity analyses lend support to the hypothesis that working memory is engaged in metacognitive monitoring, and systematic bias relies on the available information one acquires during the learning process. The different patterns of brain connectivity observed between frontal and parietal regions suggest the formation of distinct metacognitive systematic biases. These findings should enhance our understanding of the neural basis of human metacognitive systematic bias.

Data availability

The data sets generated for this study are available on request to the corresponding author.

Received: 13 January 2024; Accepted: 15 May 2024
Published online: 18 May 2024

References

1. Halamish, V. & Undorf, M. Why do judgments of learning modify memory? Evidence from identical pairs and relatedness judgments. *J. Exp. Psychol. Learn. Memory Cogn.* **49**(4), 547. https://doi.org/10.1037/xlm0001174 (2023).
2. Putnam, A. L., Deng, W. & DeSoto, K. A. Confidence ratings are better predictors of future performance than delayed judgments of learning. *Memory* **30**(5), 537–553. https://doi.org/10.1080/09658211.2022.2026973 (2022).
3. Koriat, A. & Bjork, R. A. Illusions of competence in monitoring one's knowledge during study. *J. Exp. Psychol. Learn. Memory Cogn.* **31**(2), 187. https://doi.org/10.1037/0278-7393.31.2.187 (2005).
4. Koriat, A. & Bjork, R. A. Mending metacognitive illusions: A comparison of mnemonic-based and theory-based procedures. *J. Exp. Psychol. Learn. Memory Cogn.* **32**(5), 1133. https://doi.org/10.1037/0278-7393.32.5.1133 (2006).
5. Kollmer, J., Schleinschok, K., Scheiter, K. & Eitel, A. Is drawing after learning effective for metacognitive monitoring only when supported by spatial scaffolds?. *Instr. Sci.* **48**, 569–589. https://doi.org/10.1007/s11251-020-09521-6 (2020).
6. Kubik, V., Jemstedt, A., Eshratabadi, H. M., Schwartz, B. L. & Jönsson, F. U. The underconfidence-with-practice effect in action memory: The contribution of retrieval practice to metacognitive monitoring. *Metacogn. Learn.* **17**(2), 375–398. https://doi.org/10.1007/s11409-021-09288-2 (2022).
7. Shovkova, O. & Pasichnyk, I. The illusion of thinking in metacognitive monitoring of university students. *J. Cogn. Sci.* **20**(1), 79–110. https://doi.org/10.17791/jcs.2019.20.1.79 (2019).
8. Dunlosky, J. & Nelson, T. O. Similarity between the cue for judgments of learning (JOL) and the cue for test is not the primary determinant of JOL accuracy. *J. Mem. Lang.* **36**(1), 34–49. https://doi.org/10.1006/jmla.1996.2476 (1997).
9. Nelson, T. O. & Dunlosky, J. When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The delayed-JOL effect. *Psychol. Sci.* **2**(4), 267–271. https://doi.org/10.1111/j.1467-9280.1991.tb00147.x (1991).
10. Sikström, S. & Jönsson, F. A model for stochastic drift in memory strength to account for judgments of learning. *Psychol. Rev.* **112**(4), 932. https://doi.org/10.1037/0033-295X.112.4.932 (2005).
11. Do Lam, A. T. *et al.* Monitoring the mind: The neurocognitive correlates of metamemory. *PloS One* https://doi.org/10.1371/journal.pone.0030009 (2012).
12. Morales, J., Lau, H. & Fleming, S. M. Domain-general and domain-specific patterns of activity supporting metacognition in human prefrontal cortex. *J. Neurosci.* **38**(14), 3534–3546. https://doi.org/10.1523/JNEUROSCI.2360-17.2018 (2018).

13. Qiu, L. *et al.* The neural system of metacognition accompanying decision-making in the prefrontal cortex. *PLoS Biol.* https://doi.org/10.1371/journal.pbio.2004037 (2018).
14. Kao, Y. C., Davis, E. S. & Gabrieli, J. D. Neural correlates of actual and predicted memory formation. *Nat. Neurosci.* **8**(12), 1776–1783. https://doi.org/10.1038/nn1595 (2005).
15. Frisoni, M., Di Ghionno, M., Guidotti, R., Tosoni, A. & Sestieri, C. Reconstructive nature of temporal memory for movie scenes. *Cognition* https://doi.org/10.1016/j.cognition.2020.104557 (2021).
16. Levine, L. J., Lench, H. C., Karnaze, M. M. & Carlson, S. J. Bias in predicted and remembered emotion. *Curr. Opinion Behave. Sci.* **19**, 73–77. https://doi.org/10.1016/j.cobeha.2017.10.008 (2018).
17. Persaud, K., Macias, C., Hemmer, P. & Bonawitz, E. Evaluating recall error in preschoolers: Category expectations influence episodic memory for color. *Cogn. psychol.* https://doi.org/10.1016/j.cogpsych.2020.101357 (2021).
18. Suddendorf, T. Episodic memory versus episodic foresight: Similarities and differences. *Wiley Interdiscipl. Rev. Cogn. Sci.* **1**(1), 99–107. https://doi.org/10.1002/wcs.23 (2010).
19. Carvalho, M., Cooper, A. & Marmurek, H. H. Covert retrieval yields a forward testing effect across levels of successive list similarity. *Metacogn. Learn.* **18**(3), 847–861. https://doi.org/10.1007/s11409-023-09348-9 (2023).
20. Tekin, E. & Roediger, H. L. The effect of delayed judgments of learning on retention. *Metacogn. Learn.* **16**, 407–429. https://doi.org/10.1007/s11409-021-09260-0 (2021).
21. Krasnoff, J. & Souza, A. S. I remember it now, so I'll remember it later: Working memory strength guides predictions for long-term memory performance. *Memory Cogn.* https://doi.org/10.3758/s13421-023-01514-3 (2024).
22. Jang, Y., Lee, H., Kim, Y. & Min, K. The relationship between metacognitive ability and metacognitive accuracy. *Metacogn. Learn.* **15**, 411–434. https://doi.org/10.1007/s11409-020-09232-w (2020).
23. Murphy, D. H., Huckins, S. C., Rhodes, M. G. & Castel, A. D. The effect of perceptual processing fluency and value on metacognition and remembering. *Psychonomic Bull. Rev.* https://doi.org/10.3758/s13423-021-02030-8 (2022).
24. Yang, C. *et al.* How to assess the contributions of processing fluency and beliefs to the formation of judgments of learning: Methods and pitfalls. *Metacogn. Learn.* **16**, 319–343. https://doi.org/10.1007/s11409-020-09254-4 (2021).
25. Molenberghs, P., Trautwein, F. M., Böckler, A., Singer, T. & Kanske, P. Neural correlates of metacognitive ability and of feeling confident: A large-scale fMRI study. *Soc. Cogn. Affect. Neurosci.* **11**(12), 1942–1951. https://doi.org/10.1093/scan/nsw093 (2016).
26. Kelley, T. D., McNeely, D. A., Serra, M. J. & Davis, T. Delayed judgments of learning are associated with activation of information from past experiences: A neurobiological examination. *Psychol. Sci.* **32**(1), 96–108. https://doi.org/10.1177/0956797620958004 (2021).
27. Yu, Y., Jiang, Y. & Li, F. The effect of value on judgment of learning in tradeoff learning condition: The mediating role of study time. *Metacogn. Learn.* **15**, 435–454. https://doi.org/10.1007/s11409-020-09234-8 (2020).
28. Ashburner, J. *et al.* SPM12 Manual The FIL Methods Group (and honorary members). Functional Imaging Laboratory (Functional Imaging Laboratory, Wellcome Trust Centre for Neuroimaging Institute of Neurology, UCL, 2014).
29. Yang, H. *et al.* Differential neural correlates underlie judgment of learning and subsequent memory performance. *Front. Psychol.* **6**, 1699. https://doi.org/10.3389/fpsyg.2015.01699 (2015).
30. Maniscalco, B. & Lau, H. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Conscious. Cogn.* **21**(1), 422–430. https://doi.org/10.1016/j.concog.2011.09.021 (2012).
31. Oosterhof, N. N., Connolly, A. C. & Haxby, J. V. CoSMoMVPA: multi-modal multivariate pattern analysis of neuroimaging data in Matlab/GNU Octave. *Front. Neuroinform.* **10**, 27. https://doi.org/10.3389/fninf.2016.00027 (2016).
32. Mandelkow, H., De Zwart, J. A. & Duyn, J. H. Linear discriminant analysis achieves high classification accuracy for the BOLD fMRI response to naturalistic movie stimuli. *Front. Hum. Neurosci.* **10**, 128. https://doi.org/10.3389/fnhum.2016.00128 (2016).
33. Mandelkow, H., de Zwart, J. A. & Duyn, J. H. Effects of spatial fMRI resolution on the classification of naturalistic movies. *NeuroImage* **162**, 45–55. https://doi.org/10.1016/j.neuroimage.2017.08.053 (2017).
34. Hebart, M. N., Schriever, Y., Donner, T. H. & Haynes, J. D. The relationship between perceptual decision variables and confidence in the human brain. *Cerebral Cortex* **26**(1), 118–130. https://doi.org/10.1093/cercor/bhu181 (2016).
35. Friston, K. J. *et al.* Bayesian model reduction and empirical Bayes for group (DCM) studies. *Neuroimage* **128**, 413–431. https://doi.org/10.1016/j.neuroimage.2015.11.015 (2016).
36. Fleming, S. M., Huijgen, J. & Dolan, R. J. Prefrontal contributions to metacognition in perceptual decision making. *J. Neurosci.* **32**(18), 6117–6125. https://doi.org/10.1523/JNEUROSCI.6489-11.2012 (2012).
37. Martin, A., Lane, T. J. & Hsu, T. Y. DLPFC-PPC-cTBS effects on metacognitive awareness. *Cortex* **167**, 41–50. https://doi.org/10.1016/j.cortex.2023.05.022 (2023).
38. Rounis, E., Maniscalco, B., Rothwell, J. C., Passingham, R. E. & Lau, H. Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. *Cogn. Neurosci.* **1**(3), 165–175. https://doi.org/10.1080/17588921003632529 (2010).
39. Denervaud, S. *et al.* An fMRI study of error monitoring in Montessori and traditionally-schooled children. *NPJ Sci. Learn.* **5**(1), 11. https://doi.org/10.1038/s41539-020-0069-6 (2020).
40 Morgenroth, E. *et al.* Using connectivity-based real-time fMRI neurofeedback to modulate attentional and resting state networks in people with high trait anxiety. *NeuroI. Clin.* https://doi.org/10.1016/j.nicl.2020.102191 (2020).
41. Von der Gablentz, J., Tempelmann, C., Münte, T. F. & Heldmann, M. Performance monitoring and behavioral adaptation during task switching: An fMRI study. *Neuroscience* **285**, 227–235. https://doi.org/10.1016/j.neuroscience.2014.11.024 (2015).
42. Bor, J. *et al.* Thalamus abnormalities during working memory in schizophrenia. An fMRI study. *Schizophrenia Res.* **125**(1), 49–53. https://doi.org/10.1016/j.schres.2010.10.018 (2011).
43. Guidali, G., Pisoni, A., Bolognini, N. & Papagno, C. Keeping order in the brain: The supramarginal gyrus and serial order in short-term memory. *Cortex* **119**, 89–99. https://doi.org/10.1016/j.cortex.2019.04.009 (2019).
44. Koenigs, M., Barbey, A. K., Postle, B. R. & Grafman, J. Superior parietal cortex is critical for the manipulation of information in working memory. *J. Neurosci.* **29**(47), 14980–14986. https://doi.org/10.1523/JNEUROSCI.3706-09.2009 (2009).
45. Lei, W. *et al.* Metacognition-related regions modulate the reactivity effect of confidence ratings on perceptual decision-making. *Neuropsychologia* https://doi.org/10.1016/j.neuropsychologia.2020.107502 (2020).
46. Ye, Q., Zou, F., Lau, H., Hu, Y. & Kwok, S. C. Causal evidence for mnemonic metacognition in human precuneus. *J. Neurosci.* **38**(28), 6379–6387. https://doi.org/10.1523/JNEUROSCI.0660-18.2018 (2018).
47. Martín-Luengo, B., Zinchenko, O., Dolgoarshinnaia, A. & Leminen, A. Retrospective confidence judgments: Meta-analysis of functional magnetic resonance imaging studies. *Hum. Brain Mapping* **42**(10), 3005–3022. https://doi.org/10.1002/hbm.25397 (2021).
48. Simons, J. S., Peers, P. V., Mazuz, Y. S., Berryhill, M. E. & Olson, I. R. Dissociation between memory accuracy and memory confidence following bilateral parietal lesions. *Cereb. Cortex* **20**(2), 479–485. https://doi.org/10.1093/cercor/bhp116 (2010).
49. Zheng, Y. *et al.* Diffusion property and functional connectivity of superior longitudinal fasciculus underpin human metacognition. *Neuropsychologia* https://doi.org/10.1016/j.neuropsychologia.2021.107847 (2021).
50. Hall, C. C., Ariss, L. & Todorov, A. The illusion of knowledge: When more information reduces accuracy and increases confidence. *Organ. Behav. Hum. Dec. Process.* **103**(2), 277–290. https://doi.org/10.1016/j.obhdp.2007.01.003 (2007).
51. Koriat, A., Ma'ayan, H., Sheffer, L. & Bjork, R. A. Exploring a mnemonic debiasing account of the underconfidence-with-practice effect. *J. Exp. Psychol. Learn. Memory Cogn.* **32**(3), 595. https://doi.org/10.1037/0278-7393.32.3.595 (2006).

52 Smith, R. *et al.* The role of medial prefrontal cortex in the working memory maintenance of one's own emotional responses. *Sci. Rep.* https://doi.org/10.1038/s41598-018-21896-8 (2018).
53. Xu, X. *et al.* Disorder-and cognitive demand-specific neurofunctional alterations during social emotional working memory in generalized anxiety disorder and major depressive disorder. *J. Affect. Disord.* **308**, 98–105. https://doi.org/10.1016/j.jad.2022.04.023 (2022).
54. Vaccaro, A. G. & Fleming, S. M. Thinking about thinking: A coordinate-based meta-analysis of neuroimaging studies of metacognitive judgements. *Brain Neurosci. Adv.* https://doi.org/10.1177/2398212818810591 (2018).
55. Ni, Y., Su, J., Wang, S. & Wan, X. Association with uncertainty monitoring, not value comparison in ventromedial prefrontal cortex during value-based decisions. *Value Comp. Ventromedial Prefrontal Cortex During Value-Based Dec.* https://doi.org/10.2139/ssrn.3155882 (2018).

Author contributions
Yingjie Jiang and Xiaojing Zhang designed research, Peiyao Cong, Yingjie Jiang, Yiting Long, and Xiaojing Zhang performed research, Peiyao Cong analyzed data, Peiyao Cong , Yingjie Jiang, Yiting Long, Xiaojing Zhang, and Yanlin Guo wrote the paper.

Funding
This work was supported by the National Natural Science Foundation of China, grant number: 32271095; The Natural Science Foundation of Jilin Province, grant number: 20230101149JC.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to Y.J.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.