Behavioral/Cognitive

Ventromedial Prefrontal Cortex Drives the Prioritization of Self-Associated Stimuli in Working Memory

Shouhang Yin,[1,2] Taiyong Bi,[3] Antao Chen,[2] and Tobias Egner[4]
[1]School of Mathematics and Statistics, Southwest University, Chongqing, 400715, China, [2]Key Laboratory of Cognition and Personality of the Ministry of Education, Faculty of Psychology, Southwest University, Chongqing, 400715, China, [3]Center for Mental Health Research in School of Management, Zunyi Medical University, Guizhou, 563006, China, and [4]Center for Cognitive Neuroscience, and Department of Psychology and Neuroscience, Duke University, Durham, North Carolina 27708

Humans show a pervasive bias for processing self- over other-related information, including in working memory (WM), where people prioritize the maintenance of self- (over other-) associated cues. To elucidate the neural mechanisms underlying this self-bias, we paired a self- versus other-associated spatial WM task with fMRI and transcranial direct current stimulation (tDCS) of human participants of both sexes. Maintaining self- (over other-) associated cues resulted in enhanced activity in classic WM regions (frontoparietal cortex), and in superior multivoxel pattern decoding of the cue locations from visual cortex. Moreover, ventromedial PFC (VMPFC) displayed enhanced functional connectivity with WM regions during maintenance of self-associated cues, which predicted individuals' behavioral self-prioritization effects. In a follow-up tDCS experiment, we targeted VMPFC with excitatory (anodal), inhibitory (cathodal), or sham tDCS. Cathodal tDCS eliminated the self-prioritization effect. These findings provide strong converging evidence for a causal role of VMPFC in driving self-prioritization effects in WM and provide a unique window into the interaction between social, self-referential processing and high-level cognitive control processes.

Key words: fMRI; self-prioritization; self-reference; tDCS; ventromedial PFC; working memory

Significance Statement

People have a strong tendency to attend to self-related stimuli, such as their names. This self-bias extends to the automatic prioritization of arbitrarily self-associated stimuli held in working memory. Since working memory is central to high-level cognition, this bias could influence how we make decisions. It is therefore important to understand the underlying brain mechanisms. Here, we used neuroimaging and noninvasive neurostimulation techniques to show that the source of self-bias in working memory is the ventromedial PFC, which modulates activity in frontoparietal brain regions to produce prioritized representations of self-associated stimuli in sensory cortex. This work thus reveals a brain circuit underlying the socially motivated (self-referential) biasing of high-level cognitive processing.

Introduction

People show a pervasive bias toward preferentially processing self-related information compared with other-related information. For instance, intrinsically self-related stimuli, such as one's name or face, are prioritized in long-term memory (Kesebir and Oishi, 2010), attract attention more potently (Alexopoulos et al., 2012; Liu et al., 2016), and are perceived quicker and more faithfully than other-related stimuli (Sui et al., 2012). We have recently shown that this type of self-prioritization is evident even in working memory (WM) (Yin et al., 2019), the mental workspace where information is temporarily maintained and manipulated to guide behavior (D'Esposito and Postle, 2015). When people had to keep in mind different spatial locations, they prioritized the WM maintenance of those locations where (arbitrary) self-associated cues compared with other-associated cues had been presented, although self-associated stimuli were no more likely to be probed than other-associated stimuli (Yin et al., 2019).

Understanding the processes underlying this form of social biasing of high-level cognition has important implications, as WM representations are central to decision-making and cognitive control (Gazzaley and Nobre, 2012; Boureau et al., 2015). To gain a deeper understanding of how WM representations are

Received July 9, 2020; revised Oct. 14, 2020; accepted Nov. 15, 2020.

Author contributions: S.Y., A.C., and T.E. designed research; S.Y. and T.B. performed research; S.Y., T.B., and A.C. analyzed data; S.Y., T.B., A.C., and T.E. wrote the first draft of the paper; S.Y., T.B., A.C., and T.E. edited the paper.

This work was supported by National Natural Science Foundation of China Grant 31771254, China Postdoctoral Science Foundation 2019M663425, and China National Postdoctoral Program for Innovative Talents Fellowship BX20200283.

The authors declare no competing financial interests.

Correspondence should be addressed to Antao Chen at xscat@swu.edu.cn.

https://doi.org/10.1523/JNEUROSCI.1783-20.2020

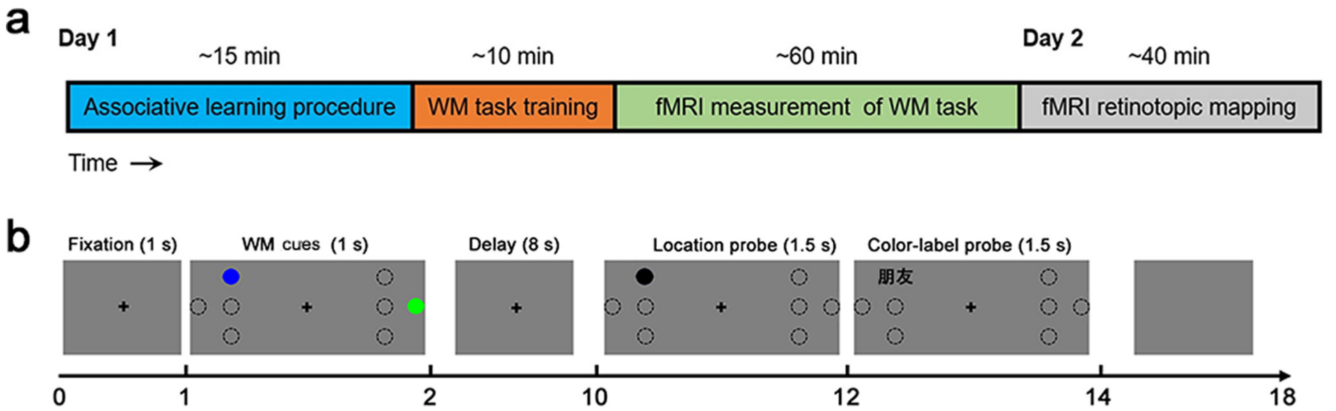

Figure 1. Task protocol and example stimuli. ***a***, The overall experiment procedure consisted of a learning phase, a training phase, the fMRI WM task, and a subsequent retinotopic mapping scan. ***b***, Example stimuli and timing of presentation of a single trial of the WM task. Participants had to remember the locations of two different color cues (previously associated with different social labels), each of which could occur in one of four locations (one cue per visual hemifield), as indicated by the dotted circle placeholders. After an 8 s delay, they responded (yes/no) to a WM probe shown at one of the locations. If the trial was a match trial, the location probe response was followed by a verbal probe for the social label associated with the color (e.g., "friend"), to which the participant had to give another yes/no response. The unit of the numbers under horizontal axis is second.

biased toward self-associated information, we paired a self- versus other-associated spatial WM task (Yin et al., 2019) with fMRI and transcranial direct current stimulation (tDCS). Specifically, participants were first trained to form associations between three colors and three persons: one with themselves, one with a best friend, and the third with a stranger. Then, they performed a delayed match-to-sample spatial WM task where they needed to memorize the locations and social labels of two color cues and then completed a recognition test. We tested two key neural predictions, derived from the literature (see below): (1) the behavioral effect of self-prioritization in WM would be mirrored by enhanced activity for, and more faithful representation of, self-associated items in brain regions supporting WM; and (2) this effect would arise from the influence on WM regions by brain areas specialized for processing self-related information.

First, a large neuroimaging literature has outlined a WM network consisting of dorsolateral PFC, the frontal eye field [FEF]), and posterior parietal cortex, including the intraparietal sulcus (IPS) and superior parietal lobule (SPL) (Baluch and Itti, 2011; Petersen and Posner, 2012). If self-associated stimuli were afforded special priority in WM, we would expect activity in these regions to be enhanced when keeping self- compared with other-associated items in mind. Moreover, the currently predominant sensory recruitment hypothesis of WM (Serences, 2016; Scimeca et al., 2018) holds that frontoparietal cortex is responsible for activating (or attending to) representations of WM items, but that those representations are maintained in, and thus decodable from, sensory cortex (Sprague et al., 2014; Rahmati et al., 2018; Cai et al., 2019; Rademaker et al., 2019). Accordingly, we expected the decoding of WM cue locations from activity patterns in visual cortex to be superior for self- than for other-associated cue locations.

Second, previous studies have consistently implicated midline structures of the ventral medial PFC (VMPFC) and the posterior cingulate cortex, key nodes of the default mode network (Raichle, 2015), when contrasting self- with other-referential processing (Qin et al., 2012; Sui et al., 2013; Yankouskaya et al., 2017). We expected to replicate this finding here in the domain of WM. Moreover, we expected that these self-referential processing regions would exhibit increased functional coupling with WM-related regions during the maintenance of self- compared with other-associated items, reflecting the hypothesized biasing of the WM network. Finally, based on fMRI results conforming to the above predictions, we performed a follow-up tDCS experiment where we targeted VMPFC in three independent groups of participants who received anodal, cathodal, or sham stimulation. If VMPFC contributed causally to the self-prioritization effect in WM, we would expect to see this effect enhanced under anodal compared with sham stimulation or diminished under cathodal compared with sham stimulation.

Materials and Methods

Participants. Thirty-four participants took part in the fMRI study. Of those 34, 2 terminated the scan prematurely, and data from 4 other participants were excluded because of excessive head motion (3 participants, >3 mm or 3 degrees) or poor WM task performance (1 participant, <80%). Another 2 participants were excluded only from the visual cortex decoding analysis because of excessive head motion during the retinotopic mapping scan (>3 mm or 3 degrees). Thus, after exclusion, 28 participants (11 females, mean age = 20.47 years, SD = 0.97 years) remained for the main fMRI data analyses, and 26 participants (10 females, mean age = 20.50 years, SD = 1.00 years) remained for the visual cortex decoding analysis. Ninety new participants were recruited for the tDCS study, and split into three groups: anode (15 females, mean age = 20.85 years, SD = 1.45 years), cathode (15 females, mean age = 21.18 years, SD = 1.61 years), and sham (15 females, mean age = 20.89 years, SD = 1.74 years). All participants were right-handed with reported normal or corrected-to-normal vision and had no known neurologic or visual disorders. Both experiments were approved by the University Human Ethics Committee of Southwest University (China). All volunteers gave informed written consent and were compensated for their participation.

Stimuli and procedure of fMRI WM task. The full timeline of the procedure of the present study is presented in Figure 1*a*. Before entering the scanner, participants partook in an associative learning procedure (Sui et al., 2012; Yin et al., 2019). Participants were initially instructed to associate one of the colors with the self, one with a named best friend, and one with an unfamiliar person for 60 s. These associations were counterbalanced across participants and subsequently used in the spatial WM task in the scanner. This approach of creating novel color-self/other associations avoids the confounding impact of familiarity on self-reference effects (Sui et al., 2012) and thus allowed us to probe self-prioritization in WM in a tightly controlled manner. Then, participants performed a color label matching task, where on each trial a circle (1.2° × 1.2°) in one of the three colors was presented above a black fixation cross at the center of a gray screen. One of three possible Chinese characters (for self, friend, or stranger, 2.4°/3.4° × 1.2°) was displayed below the fixation cross. The visual angle between the center of the colored circle or the word and the fixation cross was 3.5°. Participants had to indicate whether the color label pairing matched with the instructed association, using the index and middle fingers of the right hand on the keypad keys 1 and 2. Each trial started with a 500 ms fixation cross, followed by a 200 ms pairing probe, after which a blank screen was presented and participants had 1500 ms to press a key as quickly and accurately as possible. The presentation of the blank screen was terminated by key press or after 1500 ms, and the trial ended with a 500 ms feedback display. Each participant performed a block of 30 trials, and their accuracy had to be at least 80% to move on to the next phase of the study. The matching task served as training to make participants master the color label associations.

The fMRI task was a delayed match-to-sample spatial WM task adapted from our previous behavioral study (Yin et al., 2019). As displayed in Figure 1*b*, on a gray background, on each trial two different-colored cues (filled-in circles, subtending 1.2° × 1.2° of visual angle) were presented, one to the left and one to the right of a central fixation cross, in one of four possible locations. Figures 1 and 3 show the eight possible locations (four at each side of the visual field; bilateral symmetry). Two possible cue locations were located horizontally parallel with the fixation cross, with distances from fixation of 3.4° and 4.6°, respectively; the other two possible cue locations were above and below the cue that was horizontally in line with, and the closest to, the fixation cross, with vertical distances of 1.2°. A trial started with a 1000 ms fixation cross that remained on screen throughout the trial, followed by two colored, filled-in circles shown for 1000 ms. Participants were asked to remember the locations and social labels associated with these color cues (based on the prior learning task). Then the trial entered an 8000 ms delay period, after which the font of the fixation-cross turned bold for 300 ms (signaling the end of the delay period). A WM probe (a black filled-in circle) was then presented for 1500 ms at one of the eight possible locations, and the participants had to judge whether the probe location matched either of the two remembered cue locations, using the index and middle fingers of their right hand to indicate yes or no. The assignment of response finger to responses was counterbalanced across participants.

The WM probe presentation was terminated by the key press or after 1500 ms, after which an adjustable duration blank screen interval was presented to keep the entire target plus blank screen presentation time at 2000 ms. If the probe matched either of the two remembered locations (match trial) and the participant indicated this correctly, a label word (Self, Friend, Stranger) was presented at the probe location for 1500 ms, and participants were required to judge whether the label word matched the remembered color in this location. Probing the color label after match trials served to ensure that participants kept actively remembering the social labels associated with each color (not just the colors). Following the response, another adjustable duration blank screen was presented to keep the presentation time of label word plus blank screen at 2000 ms. On nonmatch trials, only a 2000 ms blank screen was presented. Finally, each trial ended with a (baseline) blank screen presentation of 4000 ms.

The different possible combinations of the color memory cues resulted in three trial types or pairings: Self-Friend, Self-Stranger, and Friend-Stranger. For instance, a Self-Friend trial may present the self-associated color cue in one of the left-hand locations and the friend-associated color cue in one of the right-hand locations. Each of these trial types occurred 64 times, including 16 match trials for each of the 2 items and 32 nonmatch trials. Together, there were 192 trials, including 32 self-match trials, 32 friend-match trials, 32 stranger-match trials, and 96 nonmatch trials, evenly broken down into 8 runs (each trial type or matching type occurred equally often in each run); all kinds of trials were presented in pseudorandom order.

Our study was specifically designed to facilitate decoding of the (self- or other-associated) cue locations from fMRI data in visual cortex, by always presenting one cue per visual hemifield, and by acquiring a retinotopic mapping and WM cue location localizer scan: A standard phase-encoded method developed by Sereno et al. (1995) was used to define retinotopic visual areas in which participants viewed a rotating wedge that created traveling waves of neural activity in visual cortex (2 runs). Another independent block-design localizer run was performed to localize the retinotopic area where the stimuli were presented in the WM task. In this run, to localize regions in visual cortex responsive to the visual field locations where the targets could appear, two flickering triangular checkerboards covering the edges of possible stimuli locations were presented on each side of the screen for 12 s. The run contained 14 checkerboard blocks, interleaved with blank screen blocks of 12 s.

Experimental design and statistical analysis. As detailed above, there were three possible combinations of the color memory cues: Self-Friend, Self-Stranger, and Friend-Stranger; and there were three types of location probe match response: self-match, friend-match, and stranger-match. In the behavioral analysis, our focus was the response times (RTs) of the location probe match trials. Thus, the task is a 3-level single-factor within-subjects design and a repeated-measures ANOVA was performed on the RT data. In univariate neuroimaging analyses, two effects were examined: one using a contrast to identify self-associated activation (self-contrast: Self-Friend > Friend-Stranger conditions), and the other one to delineate regions involved in WM maintenance (WM contrast: contrasting delay period activity for Self-Friend, Self-Stranger, and Friend-Stranger trials > baseline); both effects were analyzed with *t* tests, using correction for multiple comparisons. In the multivoxel pattern analysis (MVPA), trials were divided into two groups: one where the self-associated cue was presented in the left visual hemifield and the other-associated cue in the right visual hemifield (Self_L trials); and the other one corresponding to the opposite scenario (Self_R trials). For each group of trials, MVPAs were conducted on every time point of a trial to decode the four possible WM cue locations, and the decoding accuracies were compared between self- and other-associated cues using *t* tests, corrected for multiple comparisons. In the psychophysiological interaction (PPI) analysis, the VMPFC area activated in the univariate self-contrast was saved as a seed region mask, and the WM regions activated in the univariate WM univariate contrast were saved as a target region mask. The vector of the psychological variable of interest (Self-Friend > Friend-Stranger) was calculated to create the PPI term, and neural correlates of that interaction term were identified via a *t* test, corrected for multiple comparisons. We also used dynamic causal modeling (DCM) analysis to evaluate the direction of influences between VMPFC and WM regions. Rival models were evaluated statistically via Bayesian model comparison. The behavioral task in the tDCS experiment was identical to the fMRI WM task, except a reduction of the duration of delay period, and the tDCS experiment is a 3 (Group: excitatory, inhibitory, and sham; between-subjects) × 3 (self-reference: self-match, friend-match, and stranger-match; within-subjects) mixed design. All the statistical analyses were performed with SPSS version 22.0. Finally, summary behavioral and neuroimaging data from this study can be accessed at https://osf.io/jdwcr/?view_only=efdea02d46b1499d9c8db8692b175279.

fMRI acquisition. The WM task was run on a PC with an 18.5 inch monitor (1366 × 768 at 60 Hz), using E-prime software (version 2.0), and participants watched the screen through a mirror in the magnetic bore. Images were acquired with a Siemens 3T scanner (Siemens Magnetom Trio TIM), using a standard 12-channel radiofrequency head coil. An EPI sequence was used for the collection of functional WM task data, and 221 T2-weighted images were recorded per run (TR: 2000 ms; TE: 30 ms; flip angle: 85°; FOV: 224 × 224 mm^2; matrix size: 64 × 64; in-plane resolution: 3.5 × 3.5 mm^2; slice skip: 0.3 mm; 32 ascending 3-mm-thick slices). The retinotopic visual mapping and stimulus location localizer scans were performed on the next day after the WM task scan, and signals were acquired with an EPI sequence (TR: 2000 ms; TE: 30 ms; flip angle: 90°; FOV: 192 × 192 mm^2; matrix size: 64 × 64; in-plane resolution: 3.0 × 3.0 mm^2; 33 interleaved 3-mm-thick slices; no slice skip). The bottom slice was positioned at the bottom of the temporal lobe. A high-resolution 3D structural dataset (3D MPRAGE; TR: 2600 ms; TE: 3.02 ms; flip angle: 8°; resolution: 1 × 1 × 1 mm^3; 176 slices) was collected before the retinotopic visual mapping scan.

fMRI data preprocessing. Image preprocessing and analysis were conducted in Statistical Parametric Mapping toolbox (SPM12, Welcome Department of Imaging Neuroscience, Institute of Neurology, London). The first five images were discarded to achieve magnet-steady images. The imaging data were spatially realigned, and six head motion parameters were estimated for inclusion in the task models. Images were temporally realigned to the middle slice to correct for differences in slice timing. Head motion within any MRI session was <3 mm or 3 degrees for any subject. To normalize the functional images, each subject's structural brain image was coregistered to the mean functional image and was subsequently segmented. The parameters obtained in segmentation were used to normalize each subject's functional image onto the MNI space (resampling voxel size: 3 mm^3). A filter of 8 mm FWHM was used to spatially smooth the normalized data.

GLM for fMRI data. A GLM approach was used to estimate parameter values for event-related responses. Onsets of the retention period were extracted for three trial types, and the time series data were modeled for three different vectors, corresponding to Self-Friend, Self-Stranger, and Friend-Stranger conditions, respectively. Three additional regressors also modeled the respective probe epochs for these conditions to control for their influence on retention period activation estimates; another regressor modeled the blank screen stage as a no-task baseline. The design matrices also included six head movement parameters to account for any residual movement-related effect. All these vectors were convolved with the canonical HRF. A high-pass filter was implemented with a cutoff of 128 s to remove low-frequency drift from the time series.

For each subject, we defined the self-contrast between Self-Friend and Friend-Stranger to examine brain activation in relation to the self-prioritization effect, and another WM contrast between the three conditions and the blank screen baseline to characterize generic WM brain activation. These contrasts were then subjected to group-level one-sample *t* tests where participants were treated as random effects. Group analyses were conducted within a gray matter mask to reduce total search space. For the self-prioritization effect, we used a false discovery rate (FDR) to correct for multiple comparisons in the self-contrast with a voxelwise FDR-corrected threshold of $p < 0.05$ and an extent threshold of 30 voxels. This correction approach, which is more liberal than a familywise error correction, was chosen to gain greater sensitivity for detecting potential effects in regions associated with self-referential processing that, as part of the default mode network, would normally be expected to be relatively suppressed during a WM task. As the contrast of WM activity > baseline resulted in very broadly distributed activity, and we were interested in only the most activated (core WM network) regions, we subjected it to a more conservative correction method, with a voxelwise FDR-corrected threshold of $p < 0.001$ and an extent threshold of 50 voxels. To identify overlapping regions, we also performed a conjunction analysis by overlapping the two contrast maps resulting from the above analyses. To examine the activation patterns in regions showing both WM and self-prioritization effects in more detail, we extracted the β values from these regions for each condition, using the MarsBaR toolbox in SPM12.

Multivariate analysis for fMRI data. MVPAs were conducted using PRoNTo, a pattern recognition toolbox for neuroimaging (http://www.mlnl.cs.ucl.ac.uk/pronto) (Schrouff et al., 2013). Our primary MVPA was concerned with decoding the WM cue locations from visual ROIs based on the retinotopic mapping and WM location localizer data. The anatomic volume for each subject was transformed into the anterior commissure-posterior commissure space (Talairach space). Functional volumes of retinotopic mapping scans were preprocessed using BrainVoyager QX, including 3D motion correction, linear trend removal, and high-pass filtering (0.015 Hz). Head motion within any MRI session was <3 mm or 3 degrees for any subject. The functional volumes were then aligned to the anatomic volume and transformed into the anterior commissure-posterior commissure space. Next, voxels were selected for the MVPA based on their maximal responsiveness to both the retinotopic mapping visual field localizer and the WM stimulus localizer task (for details, see Stimuli and procedure). The 120 voxels (60 for each hemispheres) in primary visual cortex (V1) that displayed the highest responses (gauged via *t* statistics) to both localizers were selected, and preprocessed but unsmoothed data were used for classifier training. The left V1 voxels were trained to decode the locations of items that appeared on the right field of vision, and vice versa for the right V1. This decoding analysis was conducted on trials that contained self-associated WM cues, thus only including Self-Friend trials and Self-Stranger trials, but no Friend-Stranger trials. These trials were divided into two groups: one where the self-associated cue was presented in the left visual hemifield and the other-associated cue in the right visual hemifield (Self_L, 64 trials); and the other where the self-associated cue was presented in the right and the other-associated cue is in the left hemifield (Self_R, 64 trials). There were four possible cue locations on each side, and each location displayed 16 times in Self_L or Self_R trials. Four classification analyses were conducted: left V1 for self-associated cues, left V1 for other-associated cues, right V1 for self-associated cues, and right V1 for other-associated cues. In the present task, each trial contained 9 time points (TRs); accordingly, the data of each time point were used as samples once, and four classifications were conducted 9 times, one per time point. All decoding analyses were performed on single-subject data, with statistical reliability subsequently assessed across the sample. Classification was accomplished using a multiclass Gaussian process, and classifier sensitivity was examined using a leave-one-trial-per-group-out approach. Specifically, the classification prediction was performed 16 times, and 60 trials (15 trials for each location) were used as training data, leaving one trial for each location as the test trials. The significance of classifier performance was determined using two-tailed, one-sample *t* tests, testing against chance performance of 0.25 ($p < 0.05$ after FWE correction).

PPI and DCM analysis for fMRI data. PPI analyses were conducted using SPM12. Based on the results of GLM, the VMPFC area activated in the Self-Friend > Friend-Stranger contrast was saved as a seed region mask, and the (mostly frontoparietal) regions activated in the WM contrast were saved as a target region mask. For each subject, the exact VMPFC seed coordinate was defined using the peak voxel in the individual first-level contrast between Self-Friend and Friend-Stranger within the group mask. A sphere with a 6 mm radius was positioned at that peak of each subject, and the deconvolved time course of VMPFC activity in this ROI was extracted to serve as the physiological variable of interest.

The vector of the psychological variable of interest (Self-Friend > Friend-Stranger) was calculated to create the PPI term (the cross-product of the physiological and psychological variables). New SPMs were computed for each subject, including the interaction term, the physiological variable (i.e., the VMPFC activation time course), and the psychological variable, as well as six head movement parameters. We then identified brain regions within the WM mask where activation was predicted by the PPI term, reflecting a change in functional coupling with the VMPFC as a function of condition (self- vs other-associated). The VMPFC activity and the psychological regressors were treated as confound variables. Afterward, individuals' contrast images were entered into a group one-sample *t* test where participants were treated as random effects, and assessed for significance using an FDR-corrected threshold of $p < 0.05$.

PPI analysis cannot provide evidence concerning the direction of functional interactions between brain regions. To evaluate the direction of influences between VMPFC and WM regions, we therefore conducted a DCM analysis (Friston et al., 2003), using DCM12 implemented in SPM12. This analysis was not planned *a priori* and should therefore be considered exploratory. We focused on the key implication of the PPI results, namely, the possibility that VMPFC exerts a greater effect on WM network regions (here represented by the SPL) under more self-referential conditions. To assess this conjecture more directly, we used the most activated 100 voxels of the VMPFC and bilateral SPLs defined by the group-level self-contrast, and saved them as search masks. Then, for each subject, the peak activations within these masks from the first-level analysis were used to create 4-mm-radius-sphere volumes of interest, and the activity time series were extracted for each volumes of interest by computing the first eigenvector of all its voxels. These time courses were adjusted for movement parameters and other effects of no interest while preserving the effects of interest related to the three experimental conditions (Self-Friend, Self-Stranger, and Friend-Stranger).

These data were then used to test a series of models embodying different assumptions about the connectivity and directional influences between the VMPFC and bilateral SPLs. In all models, we assumed

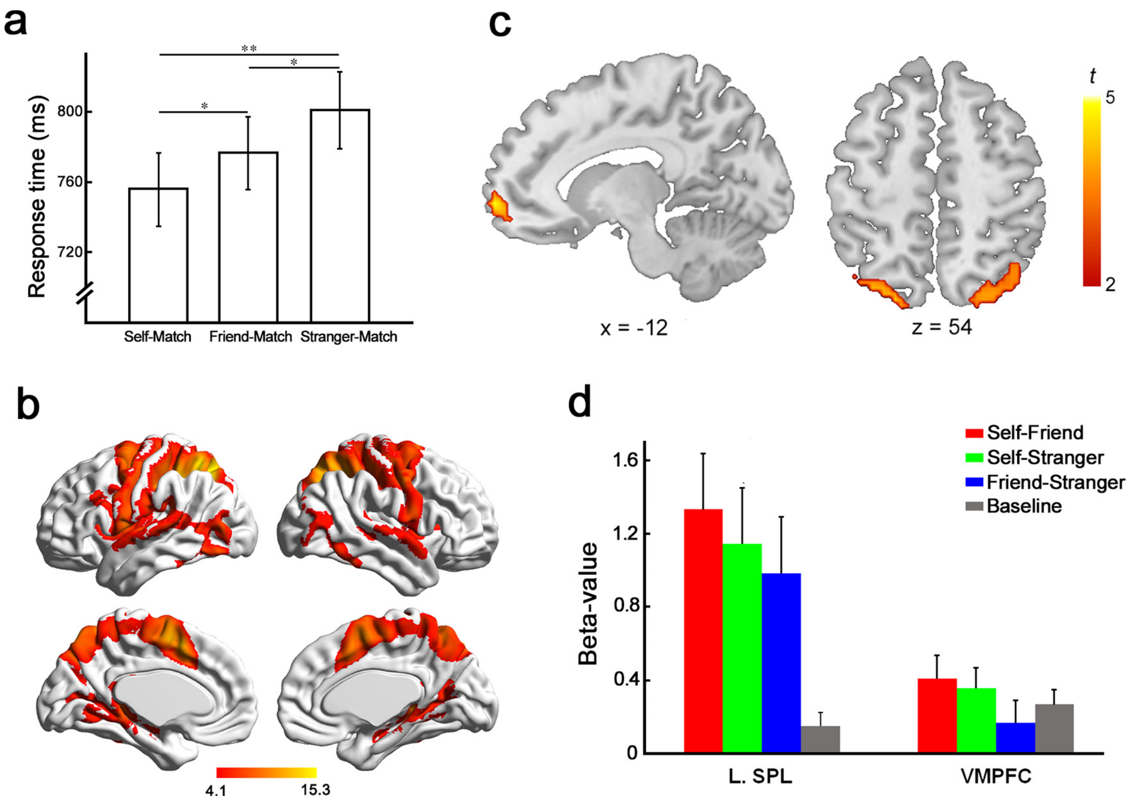

Figure 2. Behavioral and neural self-prioritization effects. ***a***, Behavioral results from the fMRI WM task replicated previous findings of a self-bias in WM. ***b***, Regions showing general involvement in WM maintenance, as defined by enhanced activity during WM delay compared with baseline, include the SMA, bilateral FEF, left IPS, bilateral SPL, bilateral precuneus, and bilateral hippocampus ($p < 0.001$, FDR-corrected). ***c***, Regions showing enhanced activation during the maintenance of more > less self-associated WM cues include both classic self-referential processing regions (VMPFC) and regions of the WM network (in particular, the SPL). ***d***, Beta values for each condition in VMPFC and left SPL $*p < 0.05$. $**p < 0.01$. Error bars indicate ± 1 SEM.

intrinsic connections within each region and extrinsic connections between left and right SPL, as well as effects of experimental conditions on each region. Here, to simplify the models, the connection pattern between VMPFC and left SPL was identical to the connection pattern between VMPFC and right SPL Thus, because of the possible connection patterns between VMPFC and bilateral SPLs, there were four context-independent intrinsic connection matrices (A-matrix): bidirectional connections between VMPFC and SPLs, connection from VMPFC to SPLs, connection from SPLs to VMPFC, and no connection between VMPFC and SPLs. Then, the possible experimental effects on the connection from VMPFC to SPLs and the connection from SPLs to VMPFC were modeled (B-matrix). There was a total of 9 models for each subject; and for each model, we derived the parameters and the free energy, which represents the log-evidence of that model. Then, we compared these models at the group level using random-effects Bayesian model selection, to identify which model had the highest probability and posterior evidence, and the most probable model was identified according to the exceedance probability (Stephan et al., 2009). The parameter values of the winning model were extracted to assess the difference among conditions using paired *t* tests.

Stimuli and procedure of tDCS task. Participants in the tDCS study performed a WM task that was identical to the fMRI WM task, except that the duration of the delay period was reduced from 8000 to 4000 ms. Before performing the WM task, participants were subjected to one of three tDCS regimens. For delivering tDCS, a DC Stimulator Plus (NeuroConn) applied a constant current of 1.5 mA for 15 min through a pair of electrodes covered in saline-soaked sponges. A 3 × 3 cm^2 forehead electrode was located at mid-distance between electrode positions Fz and Fp serving as the stimulating component, and another electrode was placed under the chin as an extracephalic reference. This electrode montage replicated prior studies demonstrating a reliable modulatory effect on hemodynamic responses in VMPFC, maximizing the unipolar stimulation of anterior VMPFC and minimizing the stimulation of other areas (Junghofer et al., 2017; Winker et al., 2018). The forehead electrode was used as the anode to produce excitatory stimulation and as a cathode to produce inhibitory stimulation (Nitsche and Paulus, 2000). Sham stimulation was performed with a current that started out the same as in the anode (or cathode) group but dropped to zero immediately after the initial current injection. The forehead electrode was used anode in half of sham group, and cathode in the other half. To control for possible trait differences in self-prioritization between groups, a measurement of narcissism was conducted for all subjects using the 16-item Narcissistic Personality Inventory (Ames et al., 2006). There was no difference between the three groups in mean RT, mean accuracy, gender, age, and narcissism score. The 16-item Narcissistic Personality Inventory measurement, associative learning procedure, and practice of WM task were performed before the stimulation, and the main WM task was performed immediately after the stimulation phase.

Results

Self-associated stimuli are prioritized in WM

Participants were highly accurate on this task, with mean accuracies for the location probe and label probe response being 96% and 95%, respectively. Since all participants' mean accuracy was higher than 95%, we did not analyze the accuracy data further. Sorted by the type of location probe match response (self-match, friend-match, and stranger-match), RT data were analyzed as a 3-level single-factor within-subjects design. Only correct responses with RTs >200 ms and within 2.5 SDs from the subject-specific mean (for each condition) were used for analysis, eliminating <1% of trials overall. These trial exclusion criteria were also applied in the subsequent tDCS study. A repeated-measures one-way ANOVA on mean RTs of location probe match trials showed a significant main effect ($F_{(2,54)}$ = 8.72, $p = 0.0005$, η^2 = 0.24, see Fig. 2*a*), with faster responses to self-

Table 1. Activated brain regions in the GLM analysis

				Peak MNI		
Contrast	Region	Cluster size	Peak *t* value	*x*	*y*	*z*
Self-Friend > Friend-Stranger	VMPFC	126	5.91	−12	66	−3
	L IFG	76	6.24	−36	27	−9
	L SPL	75	5.31	−27	−72	57
	R SPL	140	4.87	24	−75	60
Self-Friend & Self-Stranger & Friend-Stranger > blank	SMA	294	14.30	−3	6	54
	L FEF	133	12.61	−36	−3	63
	R FEF	53	9.88	30	0	66
	R hippocampus	53	11.49	21	−39	3
	L hippocampus	49	10.31	−18	−42	3
	L IPS	1037	15.30	−36	−42	42
	L precuneus	SC	15.05	−15	−72	54
	L SPL	SC	15.10	−27	−60	54
	R precuneus	735	12.82	27	−69	54
	R SPL	SC	12.18	30	−60	54

IFG, Inferior frontal gyrus; SC, same cluster.

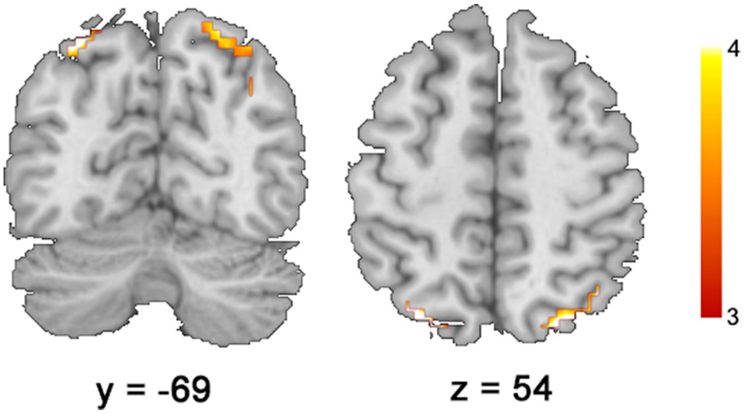

Figure 3. Regions identified by the conjunction GLM analysis. Results showed that left SPL (peak at −27, −72, 57, 32 voxels) and right SPL (peak at 24, −72, 57, 86 voxels) exhibited activation in both the self-referential processing contrast (Self-Friend > Friend-Stranger) and the WM delay period contrast (delay activity > baseline).

match trials (755.76 ± 110.00 ms) than to friend-match trials (776.47 ± 108.82 ms) ($t_{(27)}$ = 2.36, $p = 0.026$) and to stranger-match trials (800.81 ± 115.24 ms) ($t_{(27)}$ = 3.56, $p = 0.001$), as well as faster responses for friend-match trials than stranger-match trials ($t_{(27)}$ = 2.29, $p = 0.030$). These results successfully replicated those of our previous study (Yin et al., 2019), documenting the prioritization of self-associated stimuli in WM.

In the following, we test specific hypotheses of the brain mechanisms mediating this self-prioritization effect. We begin with our first prediction, that the behavioral effect of self-prioritization in WM would be mirrored by enhanced activation for self-associated items in WM regions (in addition to self-referential processing regions), and in more faithful WM representation of the location of self-associated items in visual cortex.

Enhanced activation during WM maintenance of self-associated stimuli

We began by characterizing regions involved in WM maintenance, and then assessed their activity profiles as a function of self- versus other-related item maintenance. The different possible combinations of the two memory items resulted in three trial types or pairings: Self-Friend, Self-Stranger, and Friend-Stranger. We therefore created a GLM with seven variables, three coding for the delay period for each trial type (our main task phase of interest), three coding for the location probe phase for each trial type, and one coding for the blank screen stage (baseline). To assess general involvement in WM maintenance, we initially contrasted delay period activity (collapsed across conditions) with the blank screen baseline phase (neither of these conditions displayed on-screen stimuli).

Maintaining WM representations evoked significant activity increases in the supplementary motor area (SMA), bilateral FEF, left IPS, bilateral SPL, bilateral precuneus, and bilateral hippocampus ($p < 0.001$, FDR-corrected; for more details, see Fig. 2*b*; Table 1). We next tested whether the prioritization of self-associated items in WM observed in behavior was reflected in activity levels of in WM and self-referential processing regions. To test this hypothesis, we contrasted the condition associated with the most self-referential processing (the Self-Friend condition) with that associated with the least self-referential processing (the Friend-Stranger condition). In contrasting the retention period activity between these two conditions, we found that compared with the Friend-Stranger trials, Self-Friend trials displayed greater activation in the left inferior frontal gyrus, VMPFC, and bilateral SPL ($p < 0.05$, FDR whole-brain-corrected; for more details, see Fig. 2*c*; Table 1). Thus, we observed enhanced activity for maintaining self-associated items in WM in both classic self-referential processing regions (VMPFC) and regions of the WM network (in particular, the SPL). A conjunction analysis formally confirmed the overlap between the self-referential processing effect and WM maintenance related activation in bilateral SPL (Fig. 3).

For illustrative purposes, we extracted the β values for each condition from the VMPFC and SPL regions defined by the above-reported contrast, and plotted them in Figure 2*d*. In addition to recapitulating the results of the ROI-defining contrast (i.e., greater activity in Self-Friend compared Friend-Stranger trials), these regions also displayed greater activity in the Self-Stranger compared with the Friend-Stranger conditions (VMPFC: $t_{(27)}$ = 3.55, $p = 0.001$; left SPL: $t_{(27)}$ = 2.25, $p = 0.033$; the results were equivalent in right SPL), a contrast that is orthogonal to the ROI definition (avoiding circularity). Furthermore, as expected from the WM delay period analysis above, the left SPL exhibited significantly enhanced delay period activity (over baseline) for all three trial types (all p values < 0.01). VMPFC activity during WM is generally much less pronounced than that in SPL (Fig. 2*d*). This is expected, as the VMPFC, as part of the default mode network, typically exhibits relatively suppressed activation during cognitively demanding tasks like the current one. Importantly, VMPFC shows the greatest release from this relative suppression during the conditions involving the WM maintenance of self-associated cues.

In sum, these results showed that, in addition to standard WM effects, parietal cortex also displayed a modulation of delay period activity by self-relevance, which was accompanied by typical effects of self-associated items on activity in VMPFC. These findings support one aspect of our first prediction (i.e., greater mean activity in WM regions when maintaining self-associated stimuli). We next tested the second aspect, namely, that memoranda of self-associated stimuli should be represented more faithfully than those of other-related stimuli, as assessed by decoding success of WM cue locations from delay period fMRI data.

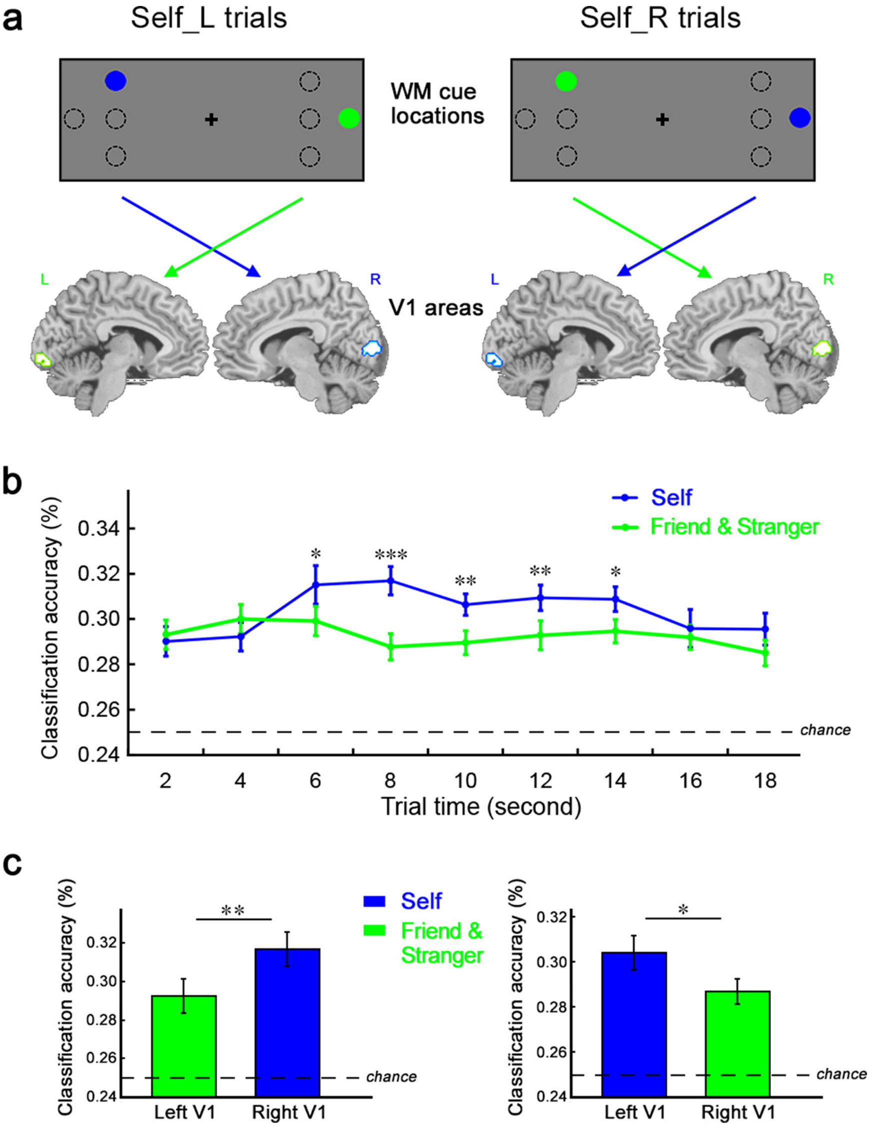

Figure 4. Decoding of self- versus other-associated WM cue locations from early visual cortex. ***a***, Examples of self- and other-associated WM cues and V1 areas from a single participant. Left, The case where a self-associated WM cue is presented in the left visual hemifield and an other-associated cue is presented in the right visual hemifield. Right, The opposite case. ***b***, Decoding performance of self- and other-associated WM cues displayed as a function of time point. For each time point, the classification accuracies of self- (Self_R and Self_L trials) and other-associated cues (Friend and Stranger trials) were averaged. The classification accuracy for self-associated cues was significantly higher than for other-associated cues at the third, fourth, fifth, sixth, and seventh time point (6-14 s after cue). Dashed line indicates the chance level (25%). ***c***, Decoding performance of simultaneously maintained self-associated and other-associated cue locations (averaged over time points 4-6). Left, Results of the self_L trials. Right, Results of the self_R trials. Vertical axis represents the mean classification accuracy. Dashed line indicates the chance level (25%). $*p < 0.05$. $**p < 0.01$. $***p < 0.001$. Error bars indicate ± 1 SEM.

Enhanced WM representation of self-associated stimuli in visual cortex

According to the "sensory recruitment" view of WM, memoranda should be maintained in relevant sensory cortex, which for the current cue items/locations would be topographically organized, early visual areas. We would not expect to observe mean (mass-univariate) activity differences between cue conditions, as we are not comparing items for which early visual cortex has differential, selective preferences. Rather, in line with previous studies, we reasoned that we should be able to decode the locations of cues held in WM from variation in multivoxel activity patterns using MVPA of activity in retinotopically organized visual areas (Sprague et al., 2014; Rahmati et al., 2018; Cai et al., 2019). Importantly, assessing the representations of WM memoranda in visual cortex allowed us to test the second aspect of our first prediction, namely, that the prioritization of self-associated information in WM should be reflected in enhanced neural representation of self-associated locations. To this end, we probed whether the neural classification of self-associated WM cue locations would display higher accuracy than that of others-associated WM cue locations.

Recall that, in the present task, there were four different possible item locations in each visual hemifield (Fig. 4; see Materials and Methods). To define visual areas with reliable retinotopy and sensitivity to stimulation at the WM cue locations, we ran a standard retinotopic localizer (Sereno et al., 1995) and a WM probe location localizer (see Materials and Methods). The intersection of visual areas identified by these localizers corresponded to left and right V1, and we used voxels within this mask for MVPA. To directly compare the neural representation of self-associated locations and other- (i.e., friend- or stranger-) associated locations, we only used the trials that involving the self-associated WM cue, and divided these trials into two categories: Self_L and Self_R trials (see Materials and Methods).

We then trained classifiers on data from left and right V1 at each time point of the WM task trials to decode which of the four possible locations in the contralateral visual hemifield was occupied by the WM cue on a given trial. We ran separate classification analyses for trials where the WM cue in the contralateral

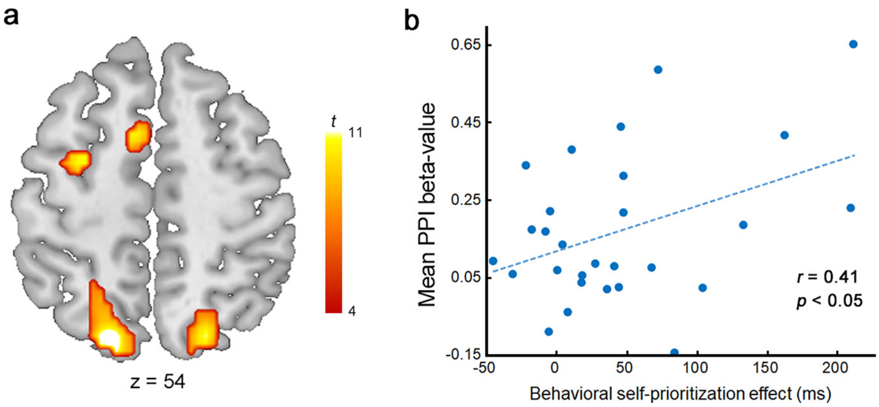

Figure 5. Functional connectivity (PPI) results. ***a***, Regions showing enhanced functional connectivity with the VMPFC (defined by the contrast shown in Fig. 1*c*) during WM maintenance of self-associated > other-associated memoranda. Enhanced coupling was observed in the SMA, left FEF, and bilateral SPL ($p < 0.05$, FDR-corrected). ***b***, A positive correlation across participants was observed between individual connection strength and behavioral self-prioritization effects. Horizontal axis represents the behavioral self-prioritization effect (defined by subtracting the self-probe's RT from stranger-probe's RT). Vertical axis represents the mean β values of the four WM regions. Scatter plot represents the line of best linear fit. Each dot represents data for a single participant.

Table 2. Brain regions exhibiting enhanced functional coupling in the PPI analysis

			Peak MNI		
Region	Cluster size	Peak *t* value	*x*	*y*	*z*
SMA	52	3.76	−6	9	54
L FEF	31	3.87	−27	−3	51
L SPL	169	4.80	−18	−75	54
R SPL	127	4.31	24	−69	60

hemifield was self-associated or other-associated (Fig. 4*a*; for more details, see Materials and Methods), resulting in a total of four classifications (left V1 for self-associated cues, left V1 for other-associated cues, right V1 for self-associated cues, right V1 for other-associated cues). For each time point, the classification accuracies of self- (Self_R and Self_L trials) and other-associated cues (Friend and Stranger trials) were averaged.

Figure 4*b* displays the decoding results, plotted as a function of time point (from 0 to 18 s). The WM cue location could be decoded at above chance levels 0.25 (all p values < 0.001, FWE-corrected) for all four classifiers. For comparison, mean mass-univariate activity in this ROI did not differentiate between the three conditions ($F_{(2,50)} = 0.33$, $p = 0.719$, $\eta^2 = 0.01$). Importantly, as shown in Figure 4*b*, paired *t* tests showed that the classification accuracy for self-associated cues was significantly higher than other-associated cues at the third ($t_{(25)} = 2.11$, $p = 0.045$), fourth ($t_{(25)} = 4.55$, $p = 0.0001$), fifth ($t_{(25)} = 3.09$, $p = 0.005$), sixth ($t_{(25)} = 3.21$, $p = 0.004$), and seventh time point ($t_{(25)} = 2.48$, $p = 0.020$). Because of hemodynamic lag, the data up until about time points 3 (6 s into the delay period) could in principle reflect differential neural responses to the cues themselves, rather than WM maintenance activity. The fact that decoding is successful, and remains superior for self-associated cue locations, over the subsequent time points (up until 14 s after cue) shows that this effect clearly extends to activity reflecting WM maintenance per se, however. We next compared the decoding performance of simultaneously maintained self-associated and other-associated cue locations using data averaged over time points 4–6 of the delay period (where decoding was most reliable). Results showed increased decoding accuracy for self-associated cue locations in contralateral visual cortex ($t_{(25)} = 3.20$, $p = 0.004$ for Self_L trials; $t_{(25)} = 2.11$, $p = 0.045$ for Self_R trials; see Fig. 4*c*). These results thus support the idea that the prioritization of self-associated stimuli in WM is reflected in enhanced neural representation of those stimuli in visual cortex.

In sum, in support of our first prediction, we observed both enhanced activation for maintaining self-associated items in frontoparietal WM regions (in particular the SPL), and more faithful representation of self-associated memoranda in visual cortex. We next turned to our second prediction, namely, that the WM self-prioritization effect arises from the influence on WM regions by brain areas specialized for processing self-related information, with the main candidate being the VMPFC region we identified above as displaying greater activation for self- than other-associated items. We first assessed this hypothesis via a functional connectivity analysis and subsequently tested it more rigorously via a tDCS experiment.

Self-associated memoranda enhance functional coupling between VMPFC and frontoparietal WM regions

To address the hypothesis that the WM network bias for self-associated cues originates with inputs from brain regions that specialize in self-related processing, we used a PPI analysis (Friston et al., 1997) to examine changes in the functional coupling (the regression slope of activation) between the VMPFC (the "seed region") and regions in the WM network (the "target regions") as a function of self-associated (Self-Friend) versus other-associated (Friend-Stranger) WM conditions. While activation in default mode regions, such as the VMPFC, typically correlates negatively with that in frontoparietal regions subserving top-down attention and WM (Fox et al., 2005; Anticevic et al., 2010; Bluhm et al., 2011; Chen et al., 2013), we here predicted the opposite (compare Spreng et al., 2010; Gerlach et al., 2011; Dixon et al., 2017). Specifically, we expected that these self-referential processing regions would exhibit a relative increase in positive functional coupling with WM-related regions during the maintenance of self-associated compared with other-related items, reflecting a biasing of the WM network. The VMPFC seed and WM search space were both defined based on the contrast results reported in the above GLM analysis (Fig. 2). We anticipated that the VMPFC would exhibit increased functional connectivity with WM regions during the maintenance of self-compared with other-associated locations.

In line with this prediction, compared with Friend-Stranger trials, Self-Friend trials showed significantly increased functional connectivity between VMPFC and the SMA, left FEF, and bilateral SPL ($p < 0.05$, FDR-corrected; for more details, see Fig. 5*a*; Table 2). To directly relate functional coupling to behavior, for each subject, we calculated the behavioral self-prioritization effect by subtracting the self-probe's RT from stranger-probe's RT, and extracted the mean β values of the above four WM regions. Then, we conducted a Pearson correlation analysis, which showed that there was a significant positive correlation

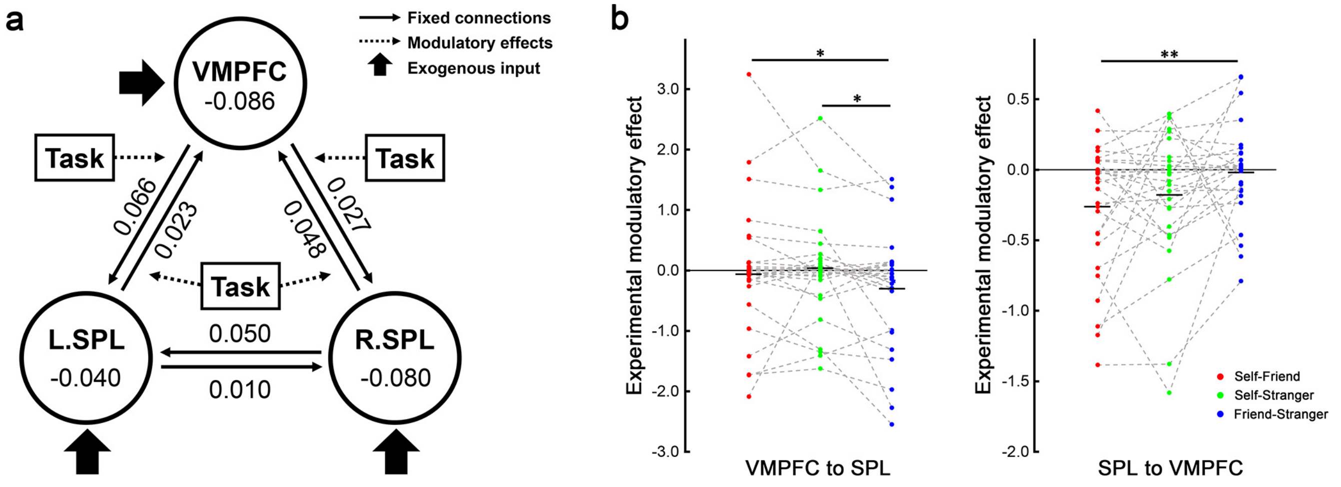

Figure 6. Winning model and parameter changes between conditions. ***a***, The structure of the winning model and the parameters of its intrinsic connections. ***b***, The modulatory effects of three experimental conditions on the connection from VMPFC to SPL (left) and the connection from SPL to VMPFC (right). Dots represent individual-participant data. Black horizontal lines indicate across-participant means. $*p < 0.05$. $**p < 0.01$.

between the mean increase in connectivity strength and the behavioral self-prioritization effect ($r = 0.41$, $p = 0.033$; see Fig. 5*b*), thus further corroborating the claim that VMPFC inputs to the WM network mediate the self-prioritization effect.

Given that PPI analysis does not convey the directionality of influence between brain regions, we followed up the above results with a DCM analysis, geared specifically at probing the interactions between VMPFC and SPL as a function of task conditions (see Materials and Methods; Fig. 6*a*). This analysis was not planned *a priori*, and the results should be considered exploratory. We estimated different models of possible influences between these regions and compared their ability to explain the data at the group level using Bayesian model selection. The winning model had an exceedance probability of 0.99, and it included nominally positive (but nonsignificant) bidirectional intrinsic coupling between VMPFC and SPLs (Fig. 6*a*) that was modulated by the experimental conditions (Fig. 6*b*; Table 3). The modulatory effect of task on all three regions' activity was more positive in the Self-Friend than in the Friend-Stranger condition ($t_{(27)} = 4.03$, $p = 0.0004$ for left SPL; $t_{(27)} = 3.30$, $p = 0.003$ for right SPL; $t_{(27)} = 3.68$, $p = 0.001$ for VMPFC). The task-dependent modulations in reciprocal influence between the VMPFC and SPL were on average inhibitory, but varied by conditions. Specifically, the influence of the VMPFC on processing in the SPL was most inhibitory in the least self-associated WM conditions, as the modulatory effect on the connection from VMPFC to SPL was more negative in Friend-Stranger than in Self-Friend ($t_{(27)} = 2.11$, $p = 0.045$) and Self-Stranger ($t_{(27)} = 2.24$, $p = 0.033$; see Fig. 6*b*) conditions. In combination with the PPI results, this could be interpreted as a release from inhibition of the VMPFC on the SPL under conditions of self-associated WM content. By contrast, the coupling from the SPL to the VMPFC became more inhibitory when moving from the less to the more self-associated WM conditions (Fig. 6*b*), as the modulatory effect was more negative in Self-Friend than in Friend-Stranger ($t_{(27)} = 2.98$, $p = 0.006$; see Fig. 6*b*) conditions. This latter finding is more difficult to reconcile with the PPI findings, but one speculative interpretation could be that the SPL's putative inhibition of the VMPFC under self-associated conditions removes the otherwise inhibitory influence of the VMPFC on the SPL. Given the *a posteriori* nature of the DCM analysis, and the complexities associated with model choices, these findings should be interpreted with caution. A future study with a design that is optimized for DCM would be required for stronger conclusions.

Table 3. Mean (SE) of the modulation parameters for experimental conditions

Condition	Self-Friend	Self-Stranger	Friend-Stranger
Regions			
L SPL	0.149 (0.050)	0.100 (0.046)	0.046 (0.061)
R SPL	0.202 (0.045)	0.162 (0.048)	0.124 (0.046)
VMPFC	0.154 (0.046)	0.125 (0.064)	−0.060 (0.039)
Connections			
L SPL to VMPFC	−0.232 (0.133)	−0.198 (0.137)	0.118 (0.149)
R SPL to VMPFC	−0.294 (0.123)	−0.148 (0.118)	−0.144 (0.100)
VMPFC to L SPL	−0.011 (0.238)	0.009 (0.203)	−0.349 (0.209)
VMPFC to R SPL	−0.029 (0.211)	−0.025 (0.166)	−0.220 (0.195)

Disrupting VMPFC with cathodal tDCS eliminates the self-prioritization effect in WM

The results of the functional connectivity analysis support the idea that VMPFC was involved in modulating activity in the WM network to favor self-associated items. However, this inference is tentative, as it is based on purely correlational data. In order to test the necessity of unperturbed VMPFC function for the self-bias in WM, we turned to the noninvasive neurostimulation technique of tDCS, which allows for drawing causal inferences. Specifically, we adopted a tDCS protocol that has recently been shown to reliably modulate VMPFC function (Junghofer et al., 2017; Winker et al., 2018) to perform excitatory (anodal), inhibitory (cathodal), and sham stimulation on this brain region in three independent groups of participants just before performing an adapted version of the above WM task (see Materials and Methods).

A 3 (group: excitatory, inhibitory, and sham; between-subjects) × 3 (self-reference: self-match, friend-match, and stranger-match; within-subjects) repeated-measures ANOVA showed no main effect of group ($F_{(2,87)} = 0.97$, $p = 0.38$, $\eta^2 = 0.02$); however, both the main effect of self-reference ($F_{(2,74)} = 27.15$, $p = 5.485 \times 10^{-11}$, $\eta^2 = 0.24$) and the interaction between group and self-reference variables ($F_{(4,174)} = 3.36$, $p = 0.011$, $\eta^2 = 0.07$) were significant, with the latter reflecting a differential impact of the stimulation protocols on self-prioritization (Fig. 7; for full behavioral data, see Table 4). To elucidate the source of this interaction, separate repeated-measures one-way

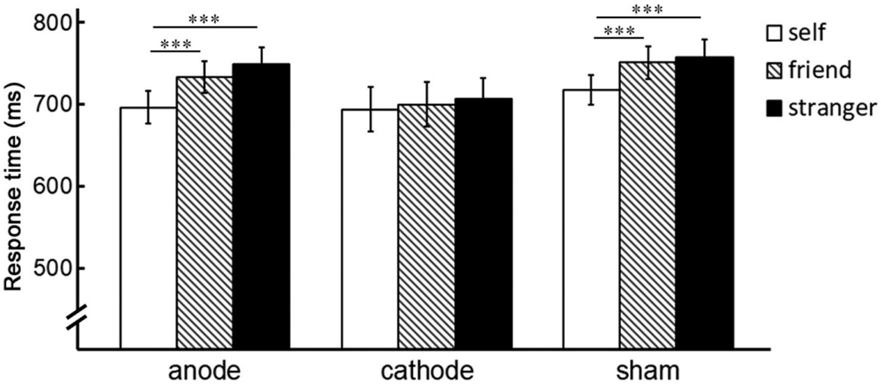

Figure 7. Behavioral results on the WM task as a function of tDCS group. A group × self-reference interaction was because the effect of self-reference was significant in the anode and sham groups but abolished in the cathode group. ***$p < 0.001$. Error bars indicate ± 1 SEM.

Table 4. Mean RT (SD) for each group/stimulation condition in the tDCS experiment

	Self	Friend	Stranger	Self-prioritization effect
Anode	696.36 (109.50)	733.63 (105.11)	749.42 (108.37)	53.06 (58.05)
Cathode	693.64 (148.93)	699.63 (148.13)	706.40 (141.19)	12.76 (47.96)
Sham	717.54 (97.59)	751.10 (108.05)	757.83 (117.76)	40.28 (51.42)

ANOVAs were conducted in each group. The main effect of self-reference was significant in the anode group ($F_{(2,58)}$ = 17.98, $p = 8.394 \times 10^{-7}$, η^2 = 0.38) and in the sham group ($F_{(2,58)}$ = 12.89, $p = 0.00002$, η^2 = 0.31), with responses to self-match trials being significantly faster than to both the friend-match trials and stranger-match trials in both groups (all p values < 0.001). However, the effect of self-reference was abolished in the cathode group ($F_{(2,58)}$ = 1.22, $p = 0.301$, η^2 = 0.04). Visual inspection of Figure 7 might lead one to suspect that this interaction effect was driven by relatively faster responses in friend and stranger trials in the cathode group. To probe this possibility, we performed three one-way between-groups ANOVAs on the RTs of self-match, friend-match, and stranger-match trials, respectively. None of these ANOVAs was significant ($F_{(2,87)}$ = 0.35, $p = 0.704$ for self-match trials; $F_{(2,87)}$ = 1.38, $p = 0.257$ for friend-match trials; $F_{(2,87)}$ = 1.50, $p = 0.228$ for stranger-match trials), indicating that the group × self-reference interaction effect was not because of a selective speedup of the friend and/or stranger conditions in the cathode group.

To directly contrast the self-prioritization effect between groups, we calculated individuals' behavioral self-prioritization effect (subtracting the self-probe's RT from stranger-probe's RT) and compared it between groups. Results showed a significant main effect of Group ($F_{(2,87)}$ = 4.59, $p = 0.013$, η^2 = 0.10), as the self-prioritization effect in cathode group (12.76 ± 47.96 ms) was significantly weaker than in the anode group (53.06 ± 58.05 ms) ($t_{(58)}$ = 2.93, $p = 0.005$) and the sham group (40.28 ± 51.42 ms) ($t_{(58)}$ = 2.14, $p = 0.036$). There was no significant enhancement of the self-prioritization effect after anodal compared with sham tDCS, possibly because of a ceiling effect. In conclusion, inhibitory (cathodal) tDCS of VMPFC removed the WM self-prioritization effect, which provides strong support for the hypothesis that VMPFC, well known for its role in self-referential processing, is the source of the self-bias observed in WM.

Discussion

The present study assessed the neural mechanisms that mediate the prioritization of self-associated information in WM. By pairing a spatial WM task involving self- and other-associated cues with fMRI, we showed that maintaining self- (vs other-) associated items robustly increased delay period activity in the VMPFC, as well as in components of the WM network, in particular the bilateral SPL. Second, using MVPA, we found that this enhanced activity when maintaining self-associated cues was accompanied by a more faithful representation (enhanced decodability) of locations corresponding to the self-associated cues in visual cortex. Third, using PPI analysis, we found that individuals' behavioral self-prioritization effect could be accounted for by increased, context-specific functional connectivity between VMPFC and WM-related regions during the maintenance of self-associated cues. DCM indicated a release of a default suppressive influence of VMPFC on SPL under self-associated WM conditions. Finally, we used tDCS to examine the causal role of the VMPFC in bringing about the WM self-prioritization effect, and found that inhibitory (cathodal) but not anodal or sham stimulation abolished the self-prioritization effect.

Our observation of enhanced WM retention period activity in VMPFC and posterior parietal cortex during the maintenance of self-associated stimuli accords well with the prior literature. The VMPFC is perhaps the most frequently implicated region in neuroimaging studies of self-referential processing (Northoff et al., 2006; Lemogne et al., 2012; Murray et al., 2012; Sui et al., 2013), whereas the SPL is a core component of the WM and dorsal (endogenous) attention networks (Baluch and Itti, 2011; Petersen and Posner, 2012; Szczepanski et al., 2013), and has been shown to support the delay period maintenance of WM items in a large number of studies (Todd and Marois, 2004; D'Esposito and Postle, 2015; Rose et al., 2016; Christophel et al., 2017). This parietal focus and an absence of strong prefrontal involvement in the current data are likely a consequence of the visuospatial nature of our WM task. Future studies would be required to generalize the current findings to more object-based WM.

In the present study, SPL activity was enhanced during the delay period per se (as in previous work), but it was further enhanced under conditions where self-associated cues had to be maintained. We interpret this activity boost during the maintenance of self-associated cues as reflecting an increased recruitment of top-down attention to support the prioritized WM status of self-associated items. While the detailed neural mechanisms of this prioritization are not yet entirely established, our speculation is concordant with recent resource-based WM accounts. In particular, it has been proposed that WM resources are flexibly (i.e., strategically) distributed among to be maintained items, and that the quality (sharpened representations, as reflected in better decodability) rather than the quantity (e.g., mean neural activity) of WM representations determines performance (Ma et al., 2014; Bays, 2015). Thus, similar to the neural and performance gains observed for retro-cued items in WM (Murray et al., 2013; Myers et al., 2015; Bays and Taylor, 2018), we speculate that the self-prioritization effect stems from a biased allocation of internal attention to the self-associated item during WM maintenance.

The notion that the increased SPL activity reflects enhanced attentional biasing of WM content is supported by our MVPA findings of more precise delay period representations of self-associated than other-associated cue locations in visual cortex. While the present paradigm was not optimized to segregate activation associated with the WM encoding versus maintenance phase, the results suggest strongly that our effects reflect WM maintenance. In particular, because of hemodynamic lag, the BOLD response associated with cue presentation/encoding would be expected to peak at ~4-6 s into the delay period. Activity related to WM maintenance would be expected to dominate the BOLD response for the subsequent 8 s (the duration of the delay period, shifted by the hemodynamic lag), that is, until ~14 s after the onset of the delay period. In line with the notion that we are capturing delay period effects, our time-resolved MVPA results revealed successful cue decoding (and an advantage for self-associated cues) throughout precisely this entire time frame, from 6 to 14 s after delay period onset (Fig. 4*b*). Especially the later parts of this phase would clearly not be expected to reflect activity related to initial cue presentation.

Prior neuroimaging studies have shown that WM contents can be decoded from multiple regions, ranging from sensory to parietal cortex and PFC (Christophel et al., 2012, 2017; Emrich et al., 2013; Sreenivasan et al., 2014). There is an ongoing debate in the literature whether (frontal and) parietal cortex is directly responsible for representing WM items or whether it supports such maintenance via top-down attentional biasing of sensory cortex (Xu, 2017; Scimeca et al., 2018). While the present study was not designed to determine the necessity of sensory cortex for maintaining WM cue, in line with the sensory recruitment hypothesis (D'Esposito and Postle, 2015; Serences, 2016), we observed clear evidence that the cued locations were indeed maintained in early visual cortex during the delay period. Most importantly for the current purpose, the decoding success for self-associated cue locations was significantly greater than that for (simultaneously presented) other-associated cue locations.

What would compel the WM network to prioritize self-associated cue locations in this manner? One can attempt to answer this question at a functional level (why?) and at a mechanistic level (how?). At the functional level, a preference for detecting, encoding, and remembering self-related information could clearly be of benefit to oneself (including at the phylogenetic time scale). Of note, this self-bias appears to be very potent and quasi-automatic: we observed this bias under conditions where we used meaningless stimuli (colored discs) that were arbitrarily associated with the self or other people, and where self-associated cue locations were no more likely to be probed than other-associated locations. Indeed, prior work has shown that this bias even persists when self-associated cues are probed less frequently than other-associated ones, that is, in situations where the self-bias is clearly not performance-conducive (Sui et al., 2014; Yin et al., 2019).

At the mechanistic level, the present study has produced compelling evidence that the neural origin of this bias lies with the VMPFC. First, as expected, the VMPFC exhibited enhanced activity under conditions of self- compared with other-associated WM maintenance, confirming its prominent role in self-referential processing (Northoff et al., 2006; Qin et al., 2012; Yankouskaya et al., 2017). Second, using PPI analysis, we found that delay periods where self-associated cues were maintained were characterized by a selective increase in functional connectivity (or a decrease in suppression, as found using DCM) between the VMPFC and regions of the WM network, in particular the SPL Third, behavioral self-prioritization effects correlated with these PPI context-specific changes in functional coupling across individuals. These results, especially in light of the prior literature implicating the VMPFC in self-referential processing, are strongly suggestive of a biasing influence from the VMPFC on the WM network when self-associated cues had to be maintained. This interpretation is also congruent with previous research reporting increased functional coupling of VMPFC with temporal regions supporting social attention in a task assessing self-bias in a perceptual matching judgment (Sui et al., 2013).

Crucially, we tested the above interpretation directly by running a tDCS experiment, adopting a stimulation protocol that has recently been validated as capable of producing distinct modulatory excitatory and inhibitory effects on VMPFC responses, as measured via fMRI (Junghofer et al., 2017; Winker et al., 2018). Whereas groups of participants receiving anodal or sham stimulation displayed the same WM self-bias effect we observed in the fMRI experiment, in the group that received cathodal (inhibitory) stimulation, the self-prioritization effect was completely abolished. This represents causal evidence for the contention that the VMPFC represents the source of the self-focused biasing effects on WM, as anticipated by the above PPI findings. However, as a caveat, it should be noted that we did not directly measure tDCS effects on neural processing in VMPFC in the present experiment. While our behavioral findings are in line with the assumption that the tDCS protocol was successful in modulating VMPFC function, this inference is part reliant on prior studies (Junghofer et al., 2017; Winker et al., 2018), and additional work is still needed to corroborate the possibility of noninvasively influencing self-referential processing in VMPFC. Of note, a within-group experimental design would provide greater sensitivity for assessing such effects.

In conclusion, the present study provides novel insights into the brain mechanisms underlying a strong bias for prioritizing the maintenance of self-associated stimuli in WM. Our behavioral, fMRI, and tDCS results provide convergent evidence for the proposal that the VMPFC biases high-level cognitive processing toward self-referential information. In particular, we posit that the VMPFC biases WM representations toward self-associated items via inputs (reflected in enhanced functional coupling) to the WM network (especially posterior parietal cortex), which in turn enhances top-down attentional modulation of sensory regions to emphasize the faithful maintenance of self- (over other-) associated items in memory. Our paradigm and findings provide a unique window into the interaction between social, self-referential processing and high-level cognitive control processes.

References

Alexopoulos T, Muller D, Ric F, Marendaz C (2012) I, me, mine: automatic attentional capture by self-related stimuli. Eur J Soc Psychol 42:770–779.

Ames DR, Rose P, Anderson CP (2006) The NPI-16 as a short measure of narcissism. J Res Pers 40:440–450.

Anticevic A, Repovs G, Shulman GL, Barch DM (2010) When less is more: TPJ and default network deactivation during encoding predicts working memory performance. Neuroimage 49:2638–2648.

Baluch F, Itti L (2011) Mechanisms of top-down attention. Trends Neurosci 34:210–224.

Bays PM (2015) Spikes not slots: noise in neural populations limits working memory. Trends Cogn Sci 19:431–438.

Bays PM, Taylor R (2018) A neural model of retrospective attention in visual working memory. Cogn Psychol 100:43–52.

Bluhm RL, Clark CR, McFarlane AC, Moores KA, Shaw ME, Lanius RA (2011) Default network connectivity during a working memory task. Hum Brain Mapp 32:1029–1035.

Boureau YL, Sokol-Hessner P, Daw ND (2015) Deciding how to decide: self-control and meta-decision making. Trends Cogn Sci 19:700–710.

Cai Y, Sheldon AD, Yu Q, Postle BR (2019) Overlapping and distinct contributions of stimulus location and of spatial context to nonspatial visual short-term memory. J Neurophysiol 121:1222–1231.

Chen AC, Oathes DJ, Chang C, Bradley T, Zhou ZW, Williams LM, Glover GH, Deisseroth K, Etkin A (2013) Causal interactions between fronto-parietal central executive and default-mode networks in humans. Proc Natl Acad Sci USA 110:19944–19949.

Christophel TB, Hebart MN, Haynes JD (2012) Decoding the contents of visual short-term memory from human visual and parietal cortex. J Neurosci 32:12983–12989.

Christophel TB, Klink PC, Spitzer B, Roelfsema PR, Haynes JD (2017) The distributed nature of working memory. Trends Cogn Sci 21:111–124.

D'Esposito M, Postle BR (2015) The cognitive neuroscience of working memory. Annu Rev Psychol 66:115–142.

Dixon ML, Andrews-Hanna JR, Spreng RN, Irving ZC, Mills C, Girn M, Christoff K (2017) Interactions between the default network and dorsal attention network vary across default subsystems, time, and cognitive states. Neuroimage 147:632–649.

Emrich SM, Riggall AC, LaRocque JJ, Postle BR (2013) Distributed patterns of activity in sensory cortex reflect the precision of multiple items maintained in visual short-term memory. J Neurosci 33:6516–6523.

Fox MD, Snyder AZ, Vincent JL, Corbetta M, Van Essen DC, Raichle ME (2005) The human brain is intrinsically organized into dynamic, anticorrelated functional networks. Proc Natl Acad Sci USA 102:9673–9678.

Friston KJ, Buechel C, Fink GR, Morris J, Rolls E, Dolan RJ (1997) Psychophysiological and modulatory interactions in neuroimaging. Neuroimage 6:218–229.

Friston KJ, Harrison L, Penny W (2003) Dynamic causal modelling. Neuroimage 19:1273–1302.

Gazzaley A, Nobre AC (2012) Top-down modulation: bridging selective attention and working memory. Trends Cogn Sci 16:129–135.

Gerlach KD, Spreng RN, Gilmore AW, Schacter DL (2011) Solving future problems: default network and executive activity associated with goal-directed mental simulations. Neuroimage 55:1816–1824.

Junghofer M, Winker C, Rehbein MA, Sabatinelli D (2017) Noninvasive stimulation of the ventromedial prefrontal cortex enhances pleasant scene processing. Cereb Cortex 27:3449–3456.

Kesebir S, Oishi S (2010) A spontaneous self-reference effect in memory: why some birthdays are harder to remember than others. Psychol Sci 21:1525–1531.

Lemogne C, Delaveau P, Freton M, Guionnet S, Fossati P (2012) Medial prefrontal cortex and the self in major depression. J Affect Disord 136:e1–e11.

Liu M, He X, Rotsthein P, Sui J (2016) Dynamically orienting your own face facilitates the automatic attraction of attention. Cogn Neurosci 7:37–44.

Ma WJ, Husain M, Bays PM (2014) Changing concepts of working memory. Nat Neurosci 17:347–356.

Murray AM, Nobre AC, Clark IA, Cravo AM, Stokes MG (2013) Attention restores discrete items to visual short-term memory. Psychol Sci 24:550–556.

Murray RJ, Schaer M, Debbané M (2012) Degrees of separation: a quantitative neuroimaging meta-analysis investigating self-specificity and shared neural activation between self- and other-reflection. Neurosci Biobehav Rev 36:1043–1059.

Myers NE, Walther L, Wallis G, Stokes MG, Nobre AC (2015) Temporal dynamics of attention during encoding versus maintenance of working memory: complementary views from event-related potentials and alpha-band oscillations. J Cogn Neurosci 27:492–508.

Nitsche MA, Paulus W (2000) Excitability changes induced in the human motor cortex by weak transcranial direct current stimulation. J Physiol 527:633–639.

Northoff G, Heinzel A, de Greck M, Bermpohl F, Dobrowolny H, Panksepp J (2006) Self-referential processing in our brain: a meta-analysis of imaging studies on the self. Neuroimage 31:440–457.

Petersen SE, Posner MI (2012) The attention system of the human brain: 20 years after. Annu Rev Neurosci 35:73–89.

Qin P, Liu Y, Shi J, Wang Y, Duncan N, Gong Q, Weng X, Northoff G (2012) Dissociation between anterior and posterior cortical regions during self-specificity and familiarity: a combined fMRI–meta-analytic study. Hum Brain Mapp 33:154–164.

Rademaker RL, Chunharas C, Serences JT (2019) Coexisting representations of sensory and mnemonic information in human visual cortex. Nat Neurosci 22:1336–1344.

Rahmati M, Saber GT, Curtis CE (2018) Population dynamics of early visual cortex during working memory. J Cogn Neurosci 30:219–233.

Raichle ME (2015) The brain's default mode network. Annu Rev Neurosci 38:433–447.

Rose NS, LaRocque JJ, Riggall AC, Gosseries O, Starrett MJ, Meyering EE, Postle BR (2016) Reactivation of latent working memories with transcranial magnetic stimulation. Science 354:1136–1139.

Schrouff J, Rosa MJ, Rondina JM, Marquand AF, Chu C, Ashburner J, Phillips C, Richiardi J, Mourão-Miranda J (2013) PRoNTo: pattern recognition for neuroimaging toolbox. Neuroinformatics 11:319–337.

Scimeca JM, Kiyonaga A, D'Esposito M (2018) Reaffirming the sensory recruitment account of working memory. Trends Cogn Sci 22:190–192.

Serences JT (2016) Neural mechanisms of information storage in visual short-term memory. Vision Res 128:53–67.

Sereno M, Dale A, Reppas J, Kwong K, Belliveau J, Brady T, Rosen B, Tootell R (1995) Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. Science 268:889–893.

Sprague TC, Ester EF, Serences JT (2014) Reconstructions of information in visual spatial working memory degrade with memory load. Curr Biol 24:2174–2180.

Spreng RN, Stevens WD, Chamberlain JP, Gilmore AW, Schacter DL (2010) Default network activity, coupled with the frontoparietal control network, supports goal-directed cognition. Neuroimage 53:303–317.

Sreenivasan KK, Curtis CE, D'Esposito M (2014) Revisiting the role of persistent neural activity during working memory. Trends Cogn Sci 18:82–89.

Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ (2009) Bayesian model selection for group studies. Neuroimage 46:1004–1017.

Sui J, He X, Humphreys GW (2012) Perceptual effects of social salience: evidence from self-prioritization effects on perceptual matching. J Exp Psychol Hum Percept Perform 38:1105–1117.

Sui J, Rotshtein P, Humphreys GW (2013) Coupling social attention to the self forms a network for personal significance. Proc Natl Acad Sci USA 110:7607–7612.

Sui J, Sun Y, Peng K, Humphreys GW (2014) The automatic and the expected self: separating self- and familiarity biases effects by manipulating stimulus probability. Atten Percept Psychophys 76:1176–1184.

Szczepanski SM, Pinsk MA, Douglas MM, Kastner S, Saalmann YB (2013) Functional and structural architecture of the human dorsal frontoparietal attention network. Proc Natl Acad Sci USA 110:15806–15811.

Todd JJ, Marois R (2004) Capacity limit of visual short-term memory in human posterior parietal cortex. Nature 428:751–754.

Winker C, Rehbein MA, Sabatinelli D, Dohn M, Maitzen J, Wolters CH, Arolt V, Junghofer M (2018) Noninvasive stimulation of the ventromedial prefrontal cortex modulates emotional face processing. Neuroimage 175:388–401.

Xu Y (2017) Reevaluating the sensory account of visual working memory storage. Trends Cogn Sci 21:794–815.

Yankouskaya A, Humphreys G, Stolte M, Stokes M, Moradi Z, Sui J (2017) An anterior–posterior axis within the ventromedial prefrontal cortex separates self and reward. Soc Cogn Affect Neurosci 12:1859–1868.

Yin S, Sui J, Chiu YC, Chen A, Egner T (2019) Automatic prioritization of self-referential stimuli in working memory. Psychol Sci 30:415–423.