Behavioral/Cognitive

Decomposing Cognitive Processes in the mPFC during Self-Thinking

Marie Levorsen,[1] **Ryuta Aoki,**[2,3] **Constantine Sedikides,**[1] **and Keise Izuma**[1,4,5]

[1]School of Psychology, University of Southampton, Southampton SO17 1BJ, United Kingdom, [2]Graduate School of Humanities, Tokyo Metropolitan University, Tokyo 192-0397, Japan, [3]Human Brain Research Center, Graduate School of Medicine, Kyoto University, Kyoto 606-8501, Japan, [4]School of Economics & Management, Kochi University of Technology, Kochi 780-8515, Japan, and [5]Research Center for Mind, Brain, and Behavior, Kochi University of Technology, Kochi 780-8515, Japan

Past cognitive neuroscience research has demonstrated that thinking about both the self and other activates the medial prefrontal cortex (mPFC), a central hub of the default mode network. The mPFC is also implicated in other cognitive processes, such as introspection and autobiographical memory, rendering elusive its exact role during thinking about the self. Specifically, it is unclear whether the same cognitive process explains the common mPFC involvement or distinct processes are responsible for the mPFC activation overlap. In this preregistered functional magnetic resonance imaging study with 35 male and female human participants, we investigated whether and to what extent mPFC activation patterns during self-reference judgment could be explained by activation patterns during the tasks of other-reference judgment, introspection, and autobiographical memory. Multivoxel pattern analysis showed that only in the mPFC were neural responses both concurrently different and similar across tasks. Furthermore, multiple regression and variance partitioning analyses indicated that each task (i.e., other-reference, introspection, and memory) uniquely and jointly explained significant variances in mPFC activation during self-reference. These findings suggest that the self-reference task engages multiple cognitive processes shared with other tasks, with the mPFC serving as a crucial hub where essential information is integrated to support judgments based on internally constructed representations.

Key words: default mode; introspection; medial prefrontal cortex; memory; multivoxel pattern analysis; self

Significance Statement

This study advances our understanding of the medial prefrontal cortex (mPFC), a central hub of the default mode network, in self-referential thinking. By using functional magnetic resonance imaging, multivoxel pattern analysis, and variance partitioning, we demonstrate that mPFC activation during self-reference judgment is explained by shared and unique contributions from other cognitive processes, including other-reference, introspection, and autobiographical memory. Importantly, the mPFC is the only region where neural responses were concurrently similar and different across these tasks, suggesting its role in integrating diverse cognitive processes. These findings highlight the mPFC's critical function in gathering and integrating information for judgments based on internal representations, shedding light on its multifaceted role in self-related cognition.

Received Dec. 17, 2024; revised April 4, 2025; accepted April 8, 2025.

Author contributions: M.L., R.A., and K.I. designed research; K.I. performed research; M.L. and K.I. analyzed data; M.L., R.A., C.S., and K.I. wrote the paper.

This research was supported by a Japan Society for Promotion of Science KAKENHI Grant Number JP19K24680 (to K.I.).

The authors declare no competing financial interests.

Correspondence should be addressed to Keise Izuma at izuma.keise@kochi-tech.ac.jp.

This paper contains supplemental material available at: https://doi.org/10.1523/JNEUROSCI.2378-24.2025

https://doi.org/10.1523/JNEUROSCI.2378-24.2025

Introduction

Thinking about the self and expressing who one is to others are fundamental aspects of human experience. The self has fascinated researchers for more than a century (James, 1890; Cooley, 1902). Reflecting this enduring interest, the intricate neural architecture of the self has been a persistent focus of inquiry (Wagner et al., 2019; Frewen et al., 2020). Using neuroimaging methods such as functional magnetic resonance imaging (fMRI), studies have established that the midline structures, the medial prefrontal cortex (mPFC), and posterior cingulate cortex (PCC) are active during the self-reference task in which individuals judge if a presented personality trait or attitudinal statement describes them (Denny et al., 2012; Murray et al., 2012).

Although the robust association between the mPFC and self-reference processing has raised the possibility that the mPFC's primary function is processing self-relevant information (Kelley et al., 2002; Northoff, 2016), the mPFC is also involved in thinking about other people (Denny et al., 2012; Murray et al., 2012). Based on these observations, some researchers (Gillihan and Farah, 2005; Legrand and Ruby, 2009) criticized the self-specific view of the mPFC, arguing that some general cognitive processes are common to self-reference and other-reference processing. For example, inferential processing and memory recall are common to both (Legrand and Ruby, 2009). In other words, mPFC activation during the self-reference task might not be related to the self specifically, but rather it is a result of general cognitive processes that take place during the self-reference task, as well as during other tasks. Indeed, the mPFC and PCC are also known to be activated by autobiographical memory (Kim, 2012; Martinelli et al., 2013) and by decision-making based on internal or subjective criteria such as moral reasoning (Nakao et al., 2012).

From a broader perspective, the mPFC and PCC are considered the core hubs of the default mode network—a network of brain regions that show heightened activation at rest (Andrews-Hanna et al., 2010). These regions are activated by a variety of tasks that depend on internally constructed representations, including not only self-reference and other-reference processing or autobiographical memory but also introspection (i.e., thinking about one's own emotional states), episodic future thinking, creativity, affective decision-making, and spatial navigation (Buckner and DiNicola, 2019; Menon, 2023). For the past two decades, researchers have attempted to identify a key common cognitive process that explains mPFC's involvement in these distinct tasks. However, these attempts are often based on univariate activation overlap (or meta-analyses), and univariate activation overlap does not constitute strong evidence for a common cognitive process across tasks (Woo et al., 2014; Levorsen et al., 2023). Thus, experimental evidence on the extent to which different tasks share a common cognitive process(es) is lacking.

Recently, fMRI studies using multivoxel pattern analysis (MVPA) and representational similarity analysis (RSA) approaches have compared patterns of activation for self-reference processing to a few other tasks. For example, Chavez et al. (2017) demonstrated that self-reference processing evoked similar activation patterns in the ventral mPFC as a positive affect (Yankouskaya et al., 2017). When comparing self-reference to other-reference, studies have demonstrated distinct patterns of activation in the mPFC (Feng et al., 2018; Courtney and Meyer, 2020; Koski et al., 2020; Parelman et al., 2022). Although these results begin to clarify scholarly understanding of affective and cognitive processes during self-thinking, the degree to which other internally focused processes (namely, introspection and memory) explain self-reference remains unknown.

In the present study, using RSA and MVPA, we aimed to test similarities and differences in neural responses between the self-reference task and three other tasks that also rely on internal representation and are known to robustly activate the mPFC. These are the other-reference, autobiographical memory (Addis et al., 2007; Summerfield et al., 2009), and introspection (Gusnard et al., 2001; Goldberg et al., 2006) tasks. Furthermore, through variance partitioning analysis (VPA), we sought to quantify the extent to which explainable variance in mPFC activation patterns during self-thinking can be attributed to activation patterns from the other three tasks.

Materials and Methods

Preregistration

We preregistered the sample size, hypotheses, participant exclusion criteria, and data analysis plan at the Open Science Framework (https://osf.io/mn9fz). Unless otherwise noted, we analyzed the data in accord with the preregistration.

Participants

The experiment was approved by the Kochi University of Technology ethics committee. Before the online autobiographical memory session, participants checked a box to indicate their consent. We obtained written consent prior to the fMRI experiment.

The final sample comprised 35 Kochi University of Technology students (8 women, 27 men), ranging in age from 18 to 22 years ($M = 19.47$; $\text{SD} = 1.08$). The sample size was based on similar previous studies (Chavez et al., 2017; Yankouskaya et al., 2017; Wen et al., 2020). We remunerated them with 2,500 Japanese yen. Participants were right-handed, had no history of psychiatric disorders, and had normal or corrected-to-normal vision. We excluded data from one additional participant due to excessive head movement (preregistered exclusion criteria of >3 mm).

Experimental procedure

The experiment consisted of two parts: (1) online autobiographical memory survey and (2) fMRI experiment. The two sessions took place on separate days, 6.97 d apart on average ($\text{SD} = 2.54$).

Online autobiographical memory session

We adapted the autobiographical memory task from Wen et al. (2020). Prior to the fMRI scan, we instructed participants to write down 15 autobiographical memories, corresponding to 1 of 15 events each. These memories should pertain to an event bound to a specific time and context that occurred >1 year ago, but after their age of 10 years. The memories ought to be clear so that participants could remember the relevant people, objects, and locations in detail.

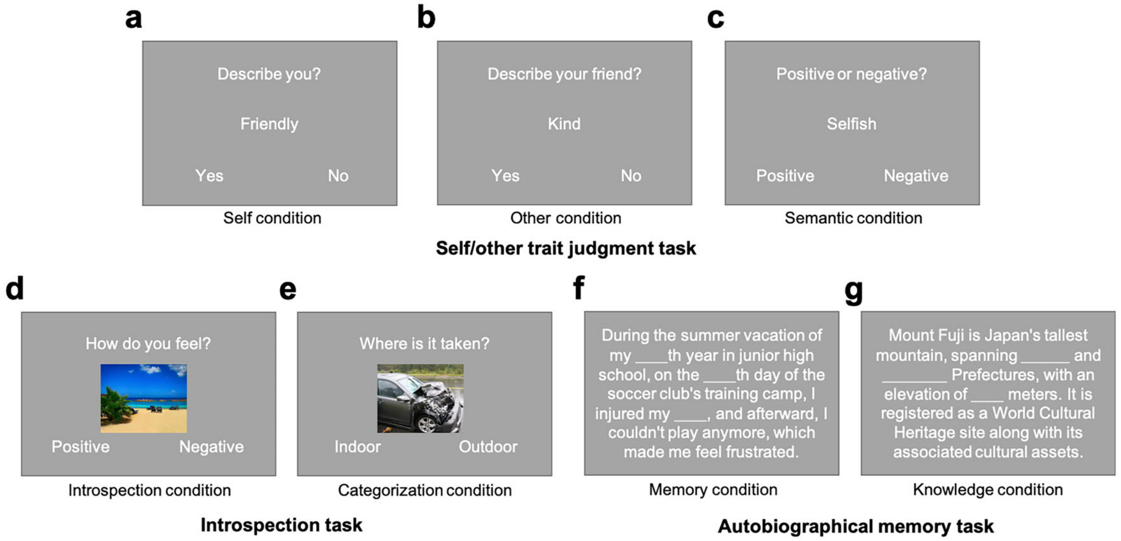
Stimuli preparation

We selected for each participant 10 of the 15 listed event memories. We used the selected memories as stimuli in the autobiographical memory condition during the fMRI experiment. We based memory selection on the amount of detail and number of characters included in each description. We matched the number of characters with stimuli in the general knowledge condition (see below). For each memory, we removed critical words and replaced them with blank underscores prior to the fMRI experiment.

The fMRI experiment

The fMRI experiment consisted of the following three tasks (Fig. 1): (1) self/other trait judgment, (2) introspection, and (3) autobiographical memory. The self/other trait judgment task had three conditions (Fig. 1*a–c*), whereas the introspection (Fig. 1*d,e*) and autobiographical memory (Fig. 1*f,g*) tasks had two conditions each. Thus, there was a total of seven conditions. Participants completed five fMRI runs, with each run lasting ∼6.5 min. Each run included two blocks of seven conditions for a total of 14 blocks. We pseudorandomized the block order within each run, so that the same task block was not presented twice in a row. At the beginning of each block, participants viewed a cue for 1 s indicating that the task that was about to commence. All text stimuli were in Japanese. We programmed all tasks in Psychtoolbox (http://psychtoolbox.org/) with MATLAB software (version 2018a; http://www.mathworks.co.uk).

Self/other trait judgment task. The stimuli comprised 40 trait adjectives from a pool of normalized trait adjectives (Anderson, 1968), which we translated into Japanese. The stimuli consisted of an equal number of positive (e.g., "honest" and "trustworthy") and negative (e.g., "mean" and



During the summer vacation of my ____th year in junior high school, on the ____th day of the soccer club's training camp, I injured my ____, and afterward, I couldn't play anymore, which made me feel frustrated.

Mount Fuji is Japan's tallest mountain, spanning ______ and ________ Prefectures, with an elevation of ____ meters. It is registered as a World Cultural Heritage site along with its associated cultural assets.

Figure 1. Examples of a trial/block for each of the seven conditions across the three tasks. The self/other trait judgment task consisted of (***a***) self-reference condition, (***b***) other-reference condition, and (***c***) semantic condition. The introspection task consisted of (***d***) introspection condition and (***e***) categorization condition. The autobiographical memory task consisted of (***f***) memory condition and (***g***) knowledge condition.

"greedy") traits. For each trial, we presented a trait in the middle of the screen. In the self-reference block (Fig. 1*a*), we asked participants to judge whether each trait describes them. In the other-reference block (Fig. 1*b*), before fMRI scanning, we asked participants to write down the name of one of their close friends on a piece of paper. During scanning, we instructed them to judge whether each trait describes this specific friend. In the semantic judgment block (Fig. 1*c*), we instructed them to judge whether each trait is positive or negative. The same 40 adjectives were used across the three tasks. We presented each trial for 2 s, followed by a 1 s fixation cross, and we presented four traits in each block (12 s per block). We randomly determined for each participant the order of traits in each of the self-reference, other-reference, and semantic conditions, but each block always included two positive and two negative words. We presented a fixation cross for 12 s before the next block.

Introspection task. We adapted the introspection task from Gusnard et al. (2001). It consisted of two conditions: introspection and categorization. We downloaded 40 picture stimuli (i.e., images of objects, animals, or sceneries) from the Open Affective Standardized Image Set (Kurdi et al., 2017). Half of the stimuli were negative and half positive. For each trial in the introspection block (Fig. 1*d*), we presented participants with an image and asked them how the image made them feel. They could respond "positive" or "negative." In the categorization block (Fig. 1*e*), we asked participants to judge whether each picture depicted a scene that was "indoors" or "outdoors." The same 40 images were used across the two tasks. For each participant, the order of images was randomly determined in each of the introspection and categorization tasks, but each block always included two positive and two negative images. We presented each image for 2 s and displayed a fixation cross for 1 s before the next image appeared (each block lasted 12 s). After an introspection/categorization block, we displayed a fixation cross for 12 s before the next block.

Autobiographical memory task. The autobiographical memory task (Wen et al., 2020) comprised two conditions: memory and knowledge. For each trial in the memory condition (Fig. 1*f*), participants encountered one of the memories they had previously listed in the online autobiographical memory session. Each memory consisted of, on average, 67.6 Japanese characters (SD = 7.25), which we matched with the length of the stimuli used in the knowledge condition. Within each memory, we replaced three critical words with blank underscores. We asked participants to recall the memory and fill in the blanks for the missing words, but do so in their mind rather than by pressing a button (i.e., we recorded no responses during this task).

In the knowledge condition (Fig. 1*g*), we presented participants with text related to general knowledge (M = 67.8 characters; SD = 8.11 characters), such as a description of a common topic (e.g., Mt. Fuji, football, and seatbelt), in which we replaced certain words with blank underscores. We instructed participants to think of appropriate words to fill in the blanks.

In both conditions, we presented each text stimulus for 14 s and followed it by a fixation cross (4–6 s). Next, we asked: "Were you recollecting a specific event?" (1, *not at all*; 5, *extremely vividly*). Participants had up to 6 s to respond. We presented a fixation cross for 10 s before the next block.

Behavioral data analysis

A one-way analysis of variance was conducted to compare reaction time (RT) and response rates across the self-reference, other-reference, and semantic judgment tasks. Given that the RT data were not normally distributed, we log-transformed them beforehand. We followed up significant effects with Bonferroni's corrected tests. All reported p values were two-sided.

We also ran a multiple regression analysis, with RT in the other-reference condition as the dependent variable and response similarity as the primary independent variable (1, same responses to the same trait; −1, different responses). Additional independent variables included participant response (1, yes; −1, no), trait valence (1, positive; −1, negative), number of characters of each word stimulus, and the interaction between participant response and trait valence. We ran the same regression analysis for RT in the self-reference condition. Due to high multicollinearity, we removed the interaction term for some participants. Also, we excluded the valence term for two participants, because it was highly positively correlated with responses, indicating strong self-enhancing and other-enhancing tendencies.

fMRI data acquisition

We acquired images using a 3.0 T Prisma Siemens MRI scanner with a 64-channel phased-array head coil. For functional imaging, we used T2*-weighted gradient-echo echo-planar imaging (EPI) sequences. We acquired 42 contiguous transaxial slices (covering almost the entire cerebrum) with a thickness of 3 mm, in an interleaved order. We acquired the images with the following parameters: time repetition, 2,500 ms; echo time, 25 ms; flip angle, 90°; field of view, 192 mm^2; and matrix, 64 × 64. Additionally, we acquired a T1-weighted structural image (with 1 mm isotropic resolution) from each participant.

fMRI data preprocessing

We carried out preprocessing and statistical analysis in SPM 12 (Welcome Department of Imaging Neuroscience), implemented in

MATLAB (MathWorks). To allow for T1 equilibration, we discarded the first four volumes before preprocessing and data analyses. We used SPM 12's preproc_fmri.m script to perform preprocessing of the fMRI data. We spatially realigned all functional images within each run to the mean using seventh-degree B-spline interpolation. We normalized the volumes to MNI space using a transformation matrix that we obtained from the EPI normalization of the first participant to the EPI template. We resampled the volumes to a voxel size of $3 \times 3 \times 3$ mm^3; i.e., we retained the original voxel size. We used the seventh-degree B-spline interpolation option for normalization. We applied spatial smoothing (of 8 mm FWHM) to the data for the whole-brain univariate analysis. To maintain fine-grained activation patterns, we did not apply smoothing to the data for RSA nor for MVPA.

Univariate fMRI analysis

General linear model (GLM)

We first ran a conventional GLM analysis, modeling separately each of the seven task blocks (i.e., conditions) with a duration of 12 s, except for the autobiographical memory and general knowledge blocks that had a duration of 14 s. The memory and knowledge tasks had a rating phase that we modeled separately as a nuisance regressor (duration = participant's response time). We also included six head motion parameters as nuisance regressors. To examine mPFC activation, we created the following six contrast images for each participant: (1) self > semantic, (2) other > semantic, (3) self > other, (4) introspection > categorization, (5) memory > knowledge, and (6) rest > semantic + knowledge + categorization. We used the last contrast to identify regions that showed increased activations during passive rest compared with externally focused tasks (Shulman et al., 1997; Gusnard et al., 2001; Wen et al., 2020). Furthermore, we created seven additional contrast images [each of the seven tasks relative to the implicit baseline (i.e., rest)]. The spmT images from these contrasts were used in the subsequent RSA and MVPA analyses (details below).

Group analysis

We conducted a second-level whole-brain group analysis for each of the contrasts. We set the statistical threshold at $p < 0.001$ voxel-wise (uncorrected) and cluster $p < 0.05$ (FWE corrected for multiple comparisons).

RSA

We conducted the RSA to test the similarity in activation patterns between the self and each of the other-reference, introspection, and memory conditions. For each participant, neural data were extracted from the spmT image of each contrast, and we computed neural representational similarity matrix (RSM; Fig. 2*a*) based on Pearson's correlation across activation patterns in each pair of conditions across the five runs. There are three model RSMs (Fig. 2*b–d*), each of which addresses the similarity between the self and (1) other, (2) introspection, and (3) memory. Given that we are interested in the similarity between the self and other, independently of similarities across the remaining conditions, we excluded from analyses the irrelevant conditions. For example, when testing the self = introspection model

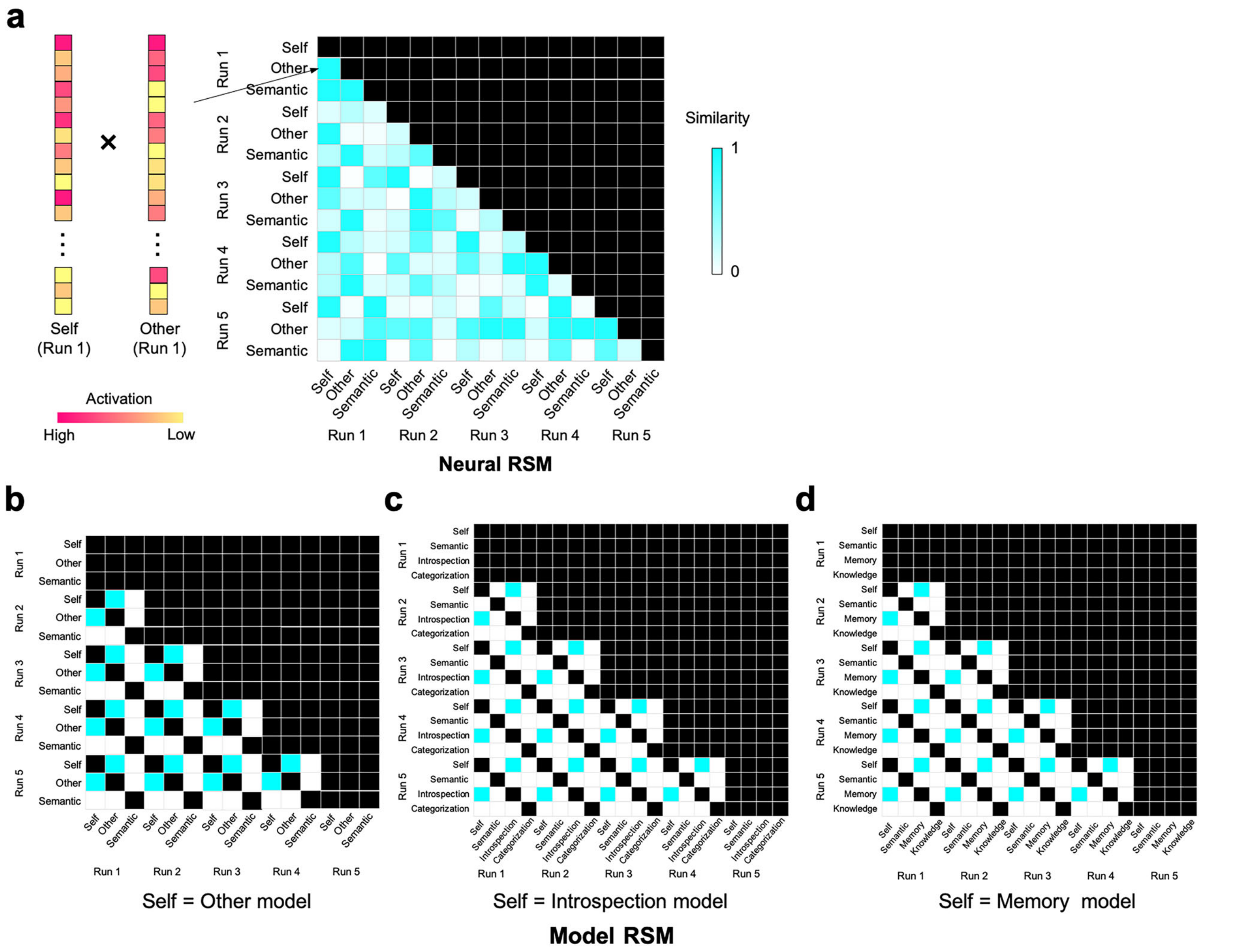


Figure 2. Schematic illustrations of RSA. ***a***, For each participant, we created a neural RSM by computing Pearson's correlations between activation patterns during two tasks across five runs. ***b***, Self = other model RSM. ***c***, Self = introspection model RSM. ***d***, Self = memory model RSM. In each neural/model RSM, we excluded cells in black from the analysis. In panels ***b–d***, cells in cyan represent 1 (similar), while cells in white represent 0 (dissimilar). We evaluated fit between the neural and each model RSMs through Kendall's tau-a (Nili et al., 2014).

(Fig. 2*c*), we excluded the other-reference, memory, and knowledge conditions so that pattern similarities involving those irrelevant conditions would not affect the results. We evaluated the fit between the neural RSM and model RSM via Kendall's tau-a for each participant (Nili et al., 2014). Activations of any two conditions within the same run are likely to be positively correlated largely due to shared physiological noises (Alink et al., 2015); as such, we excluded correlations between any pairs of conditions within the same run to the model RSM. We also excluded correlations between neural responses of the same conditions (Ritchie et al., 2017). We ran these RSAs using a searchlight approach (explained below).

Classifier-based MVPA

The above RSA tests whether activation patterns are similar between two conditions. We proceeded to conduct classifier-based MVPA to examine whether activation patterns in the two conditions were distinct. We implemented a linear support vector machine, carried out via MATLAB in combination with LIBSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm/; Wake and Izuma, 2017; Levorsen et al., 2021), with a cost parameter of $c = 1$ (default).

We used MVPA to find out if the activation patterns for the following contrasts were distinct: (1) self > semantic versus other > semantic, (2) self > semantic versus introspection > categorization, and (3) self > semantic versus memory > knowledge. For each participant, neural data were extracted from the spmT image of each of these contrasts. To evaluate classification performance, we employed a leave-one-run-out cross-validation procedure. Thus, we first left out one run in each cross-validation, and, using the data from the rest of runs, we trained a classifier that discriminates between activation patterns (e.g., self > semantic vs introspection > categorization). Subsequently, we tested the classifier performance using the data from the left-out run. We repeated this procedure five times leaving out a different run each time, and we averaged the five classification accuracy values. Like the RSA, we ran the classifier-based MVPA using a searchlight approach (below).

Searchlight analysis

We conducted the RSA and MVPA with a searchlight approach (Kriegeskorte et al., 2006). For the RSA, we extracted local patterns of neural activity from searchlights with a three-voxel radius, so that each searchlight consisted of a maximum of 123 voxels (and less on the edges of the brain). We made a neural RSM from each searchlight and computed Kendall's tau-a between neural and each of the three model RSMs (Fig. 2), which we saved for a center voxel, resulting in three correlation maps for each participant.

Similarly, for the classifier-based MVPA, we carried out MVPA within each searchlight, and we saved a classification accuracy for a center voxel, resulting in a total of three classification accuracy maps for each participant. Within each searchlight, we removed mean activity by subtracting the mean value of a searchlight sphere from values of the individual voxel so that mean activation difference across conditions could not account for MVPA results.

Group analysis

We applied smoothing before the group analysis of the RSA and MVPA outputs (with a Gaussian kernel of 4 mm FWHM). Following the smoothing, we entered Kendall's tau-a maps and classification accuracy maps into a second-level permutation-based analysis (with 5,000 permutations). We used the Statistical Non-Parametric Mapping toolbox for SPM (Nichols and Holmes, 2002). Within the preregistered mPFC region of interest (ROI), we set a statistical threshold (i.e., voxel-level) at $p < 0.005$ and a cluster-level threshold at $p < 0.05$ (FWE corrected). Outside of the mPFC, we set a statistical threshold at $p < 0.001$ and a cluster-level threshold at $p < 0.05$ (FWE corrected).

ROI analysis

We further investigated the role of the mPFC in thinking about the self by running a ROI analysis. We used Neurosynth (https://neurosynth.org/; Yarkoni et al., 2011) to define our mPFC ROI independently of our data. We downloaded an association map (thresholded at $q < 0.01$, false discovery rate corrected), which we generated from a term-based meta-analysis with the label "self-referential" (downloaded on Oct. 10, 2023). The mPFC ROI included 308 voxels. We ran the following multivariate pattern regression analyses within the ROI.

Multivariate pattern regression

The above RSA and classifier-based MVPA address neural pattern similarity and difference separately for each pair of tasks. We conducted a multivariate pattern regression analysis to compare pattern similarity across multiple tasks within the same framework. We ran a multiple regression analysis where activation patterns of the self > semantic contrast were a dependent variable, whereas those of (1) the other > semantic, (2) introspection > categorization, and (3) memory > knowledge contrasts were independent variables (Fig. 3).

As stated above, given that activation patterns of any two conditions within the same run are likely to be positively correlated due to shared physiological noise, we ran the regression analysis 20 times (i.e., all possible pairs of five runs excluding pairs from the same run) so that independent and dependent variables were always from two different runs. We averaged all outputs (i.e., beta values and adjusted R^2) across the 20 regression analyses within each participant.

Noise-ceiling model. To provide an estimate of how much systematic variation in activation patterns of the self > semantic contrast could be explained in the data given measurement noise, we included a noise-ceiling model. This model simply included the data from the self > semantic contrast as both dependent and independent variables (although they were from different runs) in the multivariate pattern regression. Thus, the only difference between the noise-ceiling model and original full model (illustrated in Fig. 3) was the inclusion of activation patterns of the self > semantic contrast as another independent variable in the noise-ceiling model.

VPA. Following the multivariate pattern regression analysis, we carried out VPA to infer the amount of unique and shared variance between three different predictors. We conducted seven multiple regression analyses: one with all three independent variables as predictors (illustrated in Fig. 3), three with different pairs of two independent variables as

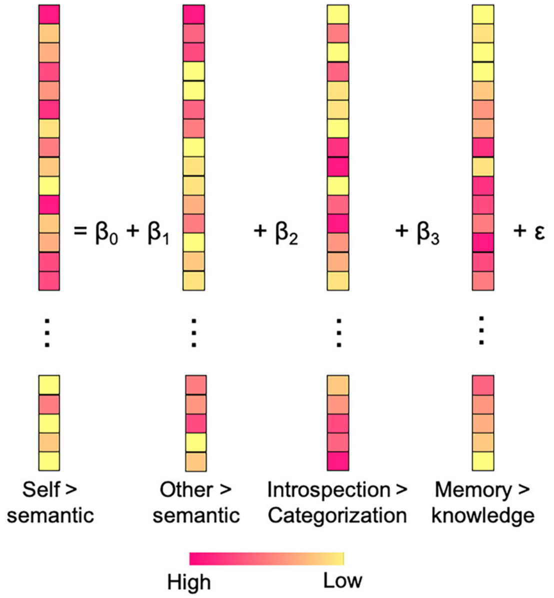


Figure 3. Multivariate pattern regression. Activation patterns of the self > semantic contrast were a dependent variable, whereas activation patterns of the other three contrast were independent variables. Independent and dependent variables were always from different runs.

predictors, and three with individual independent variables as predictors. Comparing the explained variance (R^2) of a model used alone with the explained variance when used with other models would allow us to infer the amount of unique and shared variance between different predictors.

Permutation test. To assess the significance of the findings from the multivariate pattern regression analyses and VPA, we ran permutation tests where voxels were randomly shuffled. The self > semantic contrast and other > semantic contrast have the semantic condition as a common control condition, and this common control condition is likely to bias a beta value associated with the other > semantic activation patterns to a positive direction. Thus, our permutation test randomly shuffled beta activation map of the self-reference condition (i.e., self > implicit rest contrast). We computed a randomly shuffled self > semantic contrast image (and a corresponding t-statistics map) so that the effect of the similarity in neural responses between the semantic task and each of the remaining five tasks remained intact in each permutation. We repeated this step 1,000 times to estimate null distributions. Furthermore, shuffling voxels may overly destroy spatial autocorrelation in the original data, which might bias results of the permutation test. Thus, we smoothed shuffled data via a Gaussian kernel with the standard deviation of 0.86 before conducting a multiple regression analysis (see Burt et al., 2020 for a similar approach). We selected a standard deviation of 0.86, because it produced the smallest sum of square error between the smoothness (quantified as Moran's I based on an inverse Euclidean distance matrix; Moran, 1950) of the original data and that of shuffled-and-then-smoothed data (repeated 1,000 times; we tried all standard deviation values ranging from 0 to 2.0 with an increment of 0.02).

Deviations from preregistration

We deviated from the preregistration as follows. First, we preregistered and conducted MVPA testing for pattern generalizability (i.e., cross-task classification) which, like the RSA, aims to examine the similarity in activation patterns between two conditions. However, we do not report relevant results, because they were similar to the results of the RSA described below; also, this analysis is inappropriate when testing the similarity between the self-references and other-reference conditions due to their common control condition. Second, we did not preregister the following: behavioral data analyses, RT-controlled MVPA, multivariate pattern regression, and VPAs.

Results

Behavioral results

During the self/other trait judgment conditions, participants pressed one of the two keys in almost all trials in the self (99.6%), other (99.9%), and semantic (99.9%) conditions. There was a significant difference in RT across the three conditions [$F_{(2,68)} = 19.37$; $p < 0.001$]. Pairwise t tests revealed that RTs were significantly different from each other across conditions. RTs in the self-reference condition ($M = 1.21$ s; SD = 0.18 s) were significantly longer than those in the other-reference condition ($M = 1.14$ s; SD = 0.25 s; $p_{\text{corrected}} = 0.001$) and in the semantic condition ($M = 1.08$ s; SD = 0.19 s; $p_{\text{corrected}} < 0.001$). RTs in the other condition were significantly longer than those in the semantic condition ($p_{\text{corrected}} = 0.046$).

We next examined if RTs in the other-reference condition were influenced by response similarity between the self and other, as reported in a previous study (Thornton and Mitchell, 2018). We obtained a significant effect of response similarity [$t_{(34)} = -3.80$; $p = 0.003$]. RTs were shorter when the self- and other-reference judgments for the same trait were identical (i.e., both yes or both no). Although this result suggests egocentric anchoring and adjustment in other-reference judgment, we observed a similar effect in the self-reference condition (see below). The number of characters was significantly related to RTs, meaning the more characters a word had, the slower the participant responded [$t_{(34)} = 3.83$; $p = 0.003$]. We also obtained a significant participant response × trait valence interaction [$t_{(22)} = -4.14$; $p = 0.002$]. Participants were slower to respond yes than no when judging if a negative trait described their friend, whereas they did not differ in their responses to positive traits. No other significant effect emerged.

We conducted the same regression analysis for the self-reference condition to test if RTs in the self-reference condition were influenced by response similarity between the self and other. We found significant effects of response similarity [$t_{(34)} = -2.81$; $p = 0.041$] and number of characters [$t_{(34)} = 2.97$; $p = 0.027$]. When we compared beta values for the self-reference with other-reference conditions, we observed no significant difference in the effect of response similarity on RT [$t_{(34)} = 1.30$; uncorrected $p = 0.20$], suggesting that the significant effect of the response similarity obtained in the other-reference condition might be at least partially explained by unknown stimulus features.

Consistent with prior research (Moran et al., 2006), participants were more likely to endorse a positive trait as self-descriptive and a negative trait as not self-descriptive [$t_{(34)} = 4.28$; $p < 0.001$]. However, we observed this positivity bias in the other-reference condition as well [$t_{(34)} = 8.46$; $p < 0.001$]; indeed, this bias was stronger for the other-reference than the self-reference condition, indicating that participants were more other-enhancing than self-enhancing [$t_{(34)} = 3.48$; $p = 0.001$]. These results are largely consistent with some findings, suggesting that self-enhancement is weaker for East Asian compared with Western individuals (Heine and Hamamura, 2007; but see Cai et al., 2016).

During the introspection task, participants pressed one of the two keys in almost all trials in the introspection (99.9%) and categorization (99.7%) conditions. RTs during the introspection condition ($M = 1.08$ s; SD = 0.21) were significantly faster than those during the categorization ($M = 1.14$ s; SD = 0.17) condition [$t_{(34)} = 3.79$; $p < 0.001$], likely because some pictures were ambiguous as to whether they were taken indoors or outdoors.

During the autobiographical memory task, participants successfully gave their vividness rating within the time limit of 6 s for almost all trials in the memory (99.0%) and knowledge (98.7%) conditions. Vividness ratings were significantly higher in the memory ($M = 4.37$; SD = 0.40) compared with the knowledge ($M = 2.45$; SD = 0.79) condition [$t_{(34)} = 17.73$; $p < 0.001$], testifying to the effectiveness of our memory manipulation.

fMRI results

Univariate analysis results

Replicating findings from several studies (Denny et al., 2012; Murray et al., 2012), the self > semantic contrast significantly activated the midline structure including the mPFC and PCC (Fig. 4*a*; Table S1). The other > semantic contrast activated similar regions (Fig. 4*b*; Table S2). The left temporoparietal junction (TPJ) was also commonly activated by the self and other conditions. Although there were some regions that were uniquely activated by either the self > semantic or other > semantic contrast when the self and other conditions were directly compared, the self > other contrast did not lead to any significant activation. The opposite contrast (other > self) revealed only one significant cluster in the PCC (303 voxels; $x = 6$, $y = -64$, $z = 29$).

As per prior studies (Gusnard et al., 2001; Goldberg et al., 2006), the introspection > categorization contrast significantly activated the mPFC (Fig. 4*c*). Other activated areas included the anterior cingulate cortex, temporal pole, lateral temporal cortex (LTC), and lateral occipital cortex (Table S3).

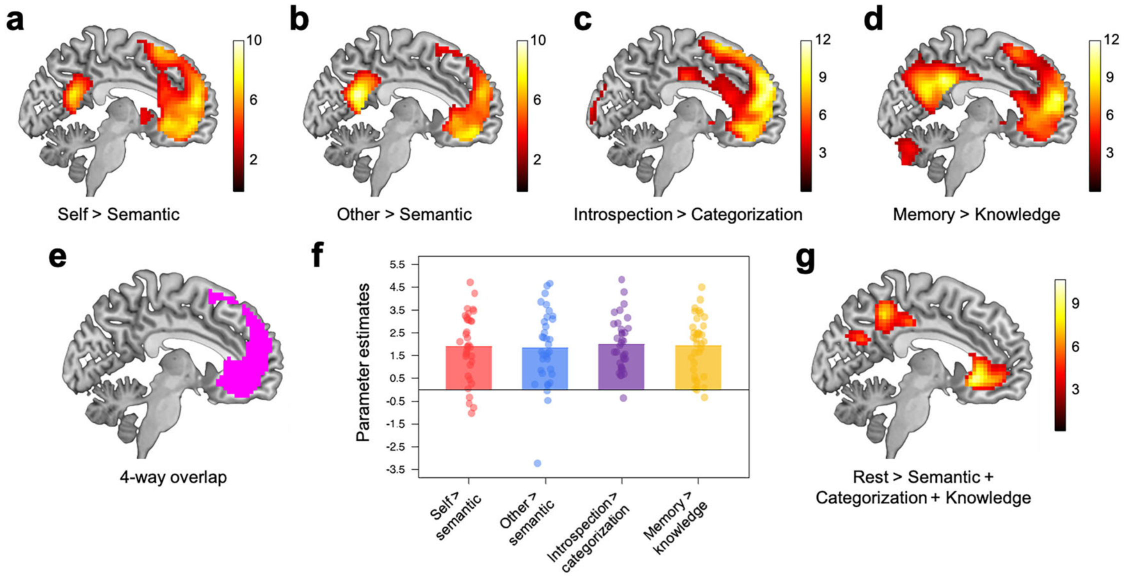


Figure 4. Sagittal slices ($x = -6$) showing results of the univariate analyses. ***a***, Areas significantly activated by the self > semantic contrast. ***b***, Areas significantly activated by the other > semantic contrast. ***c***, Areas significantly activated by the introspection > categorization contrast. ***d***, Areas significantly activated by the memory > knowledge contrast. ***e***, Areas commonly activated by all four contrasts (1,565 voxels). Only mPFC showed significant four-way overlap. For display purposes, we set statistical threshold at $p < 0.005$ and cluster $p < 0.05$ (FWE corrected). ***f***, Parameter estimates of the four contrasts within the mPFC areas commonly activated by the four contrasts (panel ***e***). ***g***, Areas significantly activated by the rest > semantic + categorization + knowledge contrast.

The memory versus knowledge contrast significantly activated regions previously implicated in autographical memory including the mPFC, PCC/precuneus, posterior inferior parietal lobule (pIPL), and LTC (Kim, 2012; Martinelli et al., 2013; Fig. 4*d*; Table S4).

Taken together, the above four contrasts all significantly activated the common region within the mPFC (1,565 voxels; Fig. 4*e*). Bilateral temporal poles were also commonly activated by all four contrasts (left $x = -36$, $y = 17$, $z = -22$, 109 voxels; right $x = 30$, $y = 14$, $z = -22$, 185 voxels). No other region was commonly activated. Yet, although the introspection > categorization contrast activated the PCC (92 voxels), it did not pass our preregistered cluster-level threshold. When we directly compared the four contrasts to each other within the commonly activated mPFC areas, no significant difference emerged [$F_{(2.38,\ 80.86)} = 0.08$; $p = 0.94$; Fig. 4*f*].

Consistent with prior findings (Shulman et al., 1997; Gusnard et al., 2001; Wen et al., 2020), the rest > semantic + categorization + knowledge contrast revealed that areas in the default mode network, including the mPFC, PCC, inferior parietal lobule (IPL), TPJ/angular gyrus (AG), and LTC, were active during rest compared with the externally focused tasks (Fig. 4*g*).

Results of RSA: are activation patterns evoked by two tasks similar?

The RSA (Fig. 2) aims to test whether the self-reference judgment evoked similar activation patterns with each of the other-reference judgment, introspection, and autobiographical memory.

The self = other model was significantly associated with a network of brain regions involved in self-reference and social cognition including the mPFC, PCC/precuneus, bilateral inferior frontal gyrus (IFG), bilateral superior temporal sulcus, and bilateral temporal pole (Fig. 5*a*). However, the other two models were associated only with the mPFC and left IFG. In particular, the self = introspection model was significantly associated with the mPFC ($x = 0$, $y = 53$, $z = 35$, 1,930 voxels; Fig. 5*b*) and left IFG (extending to the temporal pole; $x = -51$, $y = 20$, $z = 2$, 379 voxels). Further, the self = memory model was significantly associated with the mPFC ($x = -12$, $y = 44$, $z = 5$, 103 voxels; Fig. 5*c*) and left IFG ($x = -45$, $y = 26$, $z = -7$, 368 voxels).

Results of MVPA testing for pattern discriminability: are activation patterns evoked by two tasks distinguishable?

The classifier-based MVPA tested pattern discriminability with a searchlight approach. It addressed whether activation patterns evoked by different tasks were distinguishable or linked to different cognitive processes. Indeed, activation patterns evoked during the self-reference task (relative to the semantic task) were distinguishable from the other-reference task in the mPFC, PCC, and right superior temporal sulcus (extending to the temporal pole; Fig. 5*d*). These areas largely overlapped with the areas activated by the self-reference and other-reference condition relative to the semantic condition (Fig. 4*a,b*), indicating that those areas were commonly activated both by the self and other conditions compared with the semantic condition, but their activation patterns were systematically different. Given that the self and other conditions had the semantic condition as common control, the difference between the self and other conditions is likely to be underestimated in this analysis.

In contrast, activation patterns elicited by the self-reference condition were distinguishable from each of the introspection and memory conditions in a number of regions across the whole brain including the mPFC, PCC, IPL, middle temporal gyrus, and TPJ (Fig. 5*e,f*).

These results, together with the RSA results reported above, indicate that mPFC activation patterns during the self-reference judgment were similar to those elicited during each of the other-reference judgment, introspection, and autobiographical memory (Fig. 5*a–c*). Nonetheless, they were still distinguishable from activation patterns of each of the three tasks (Fig. 5*d–f*). In fact, there was one cluster within the mPFC (Fig. 5*g*; a total of 96 voxels) showing significant association/classification

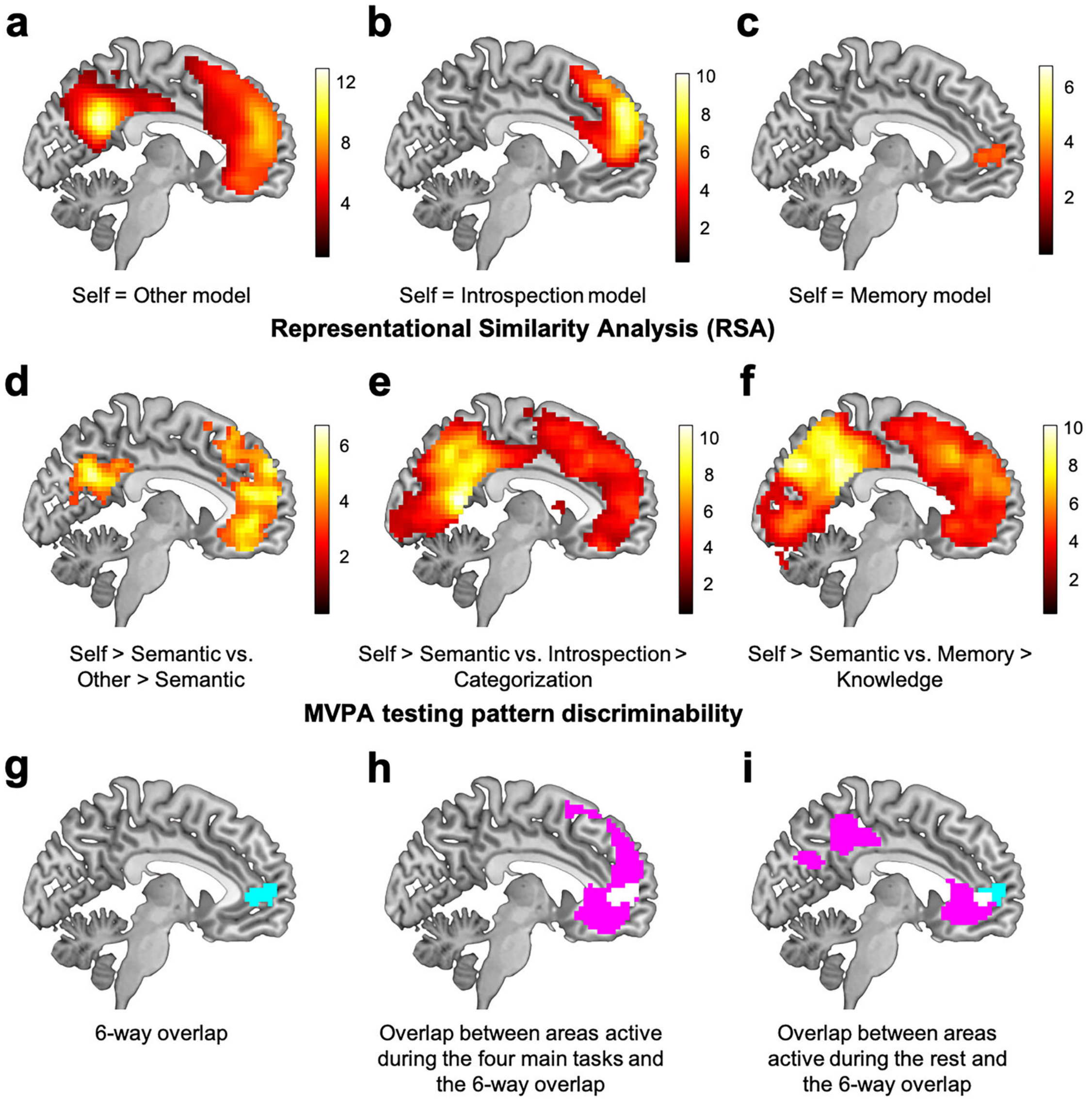

a
12
8
4
Self = Other model
b
10
8
6
4
2
Self = Introspection model
c
6
4
2
Self = Memory model
Representational Similarity Analysis (RSA)
d
6
4
2
Self > Semantic vs.
Other > Semantic
e
10
8
6
4
2
Self > Semantic vs. Introspection >
Categorization
f
10
8
6
4
2
Self > Semantic vs. Memory >
Knowledge
MVPA testing pattern discriminability
g
6-way overlap
h
Overlap between areas active
during the four main tasks and
the 6-way overlap
i
Overlap between areas
active during the rest and
the 6-way overlap

Figure 5. ***a–c***, Sagittal slices ($x = -6$) showing results from the RSA. Significant areas indicate that activation patterns of the two conditions were similar. ***d–f***, Sagittal slices ($x = -6$) showing results from the MVPA testing for pattern discriminability. Significant areas indicate that activation patterns of the two contrasts were distinguishable. For display purposes, we set statistical threshold at $p < 0.005$ and cluster $p < 0.05$ (FWE corrected). ***g***, A sagittal slice ($x = -6$) showing the mPFC area that showed six-way overlap (overlap across areas shown in panel ***a–f***). ***h***, A sagittal slice ($x = -6$) showing overlap between univariate and RSA/MVPA results. Magenta represents areas activated commonly by the four univariate contrasts (Fig. 4*e*), and white represents the six-way overlapped region depicted in panel ***g***. ***i***, A sagittal slice ($x = -6$) showing overlap (white areas) between areas activated by the rest > semantic + categorization + knowledge contrast (magenta; Fig. 4*g*) and the six-way overlapped region depicted in panel ***g*** (cyan).

accuracy in all six analyses (Fig. 5*a–f*), and the mPFC cluster was the only region that showed the six-way overlap with the cluster size larger than 20 voxels. This six-way overlap was located in the anterior part of the mPFC [Brodmann area (BA) 10] and pregenual anterior cingulate cortex (BA 32). Furthermore, this cluster entirely overlapped with the areas commonly activated by the four contrasts in the univariate analyses (Fig. 5*e*). It also showed a substantial overlap [53 out of 96 voxels (55.2%)] with the areas significantly active during the rest (Fig. 5*g*), indicating that most of the six-way overlap area (Fig. 5*e*) is located in the mPFC within the default mode network.

Results of ROI analyses

We conducted additional ROI analyses to refine the findings and run control analyses. We defined the mPFC ROI independently of our own data using Neurosynth (see Materials and Methods). This ROI analysis focused on areas within the mPFC most strongly associated with self-reference processing (Fig. 6*a*).

Does the difference in activation patterns between the self- and other-reference conditions simply reflect the difference in RTs between them?

According to our behavioral results, RTs were significantly longer for the self-reference condition compared with the other-reference condition. Thus, the difference in activation patterns between the two conditions might be explained by the difference in RTs (e.g., task difficulty). To rule out this possibility, we ran additional GLM where we categorized self-reference and other-reference task blocks into short and long RT blocks based on average RTs in each block. We modeled the other five tasks in the same way as the original GLM. Then, we ran an MVPA

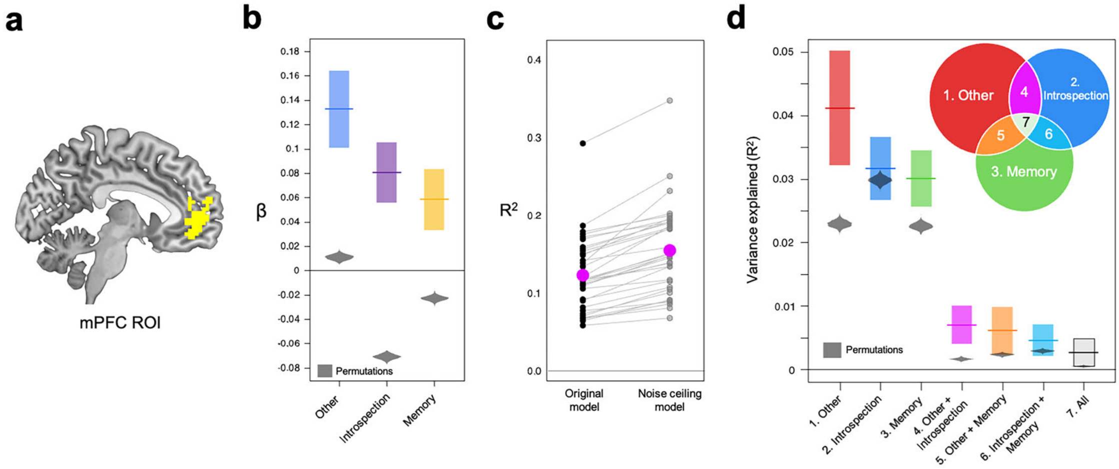


Figure 6. Results of the multivariate pattern regressions and VPA. ***a***, A sagittal slice ($x = -6$) showing the mPFC areas used in the ROI analysis. We defined the mPFC ROI with the term "self-referential" based on Neurosynth term-based meta-analysis. ***b***, Beta values from the multivariate pattern regression with activation patterns of the self > semantic contrast as a dependent variable. Colored horizontal lines indicate mean beta values, and lower/upper box limits represent 95% confidence intervals. ***c***, Adjusted R^2 from the original regression model (Fig. 3) and the noise-ceiling model. Pink circles indicate mean R^2, and black/gray circles indicate R^2 of individual subjects. ***d***, Variance in mPFC ROI activation patterns of the self-reference condition that was explained by activation patterns of the other, introspection, and memory conditions. In panels ***b*** and ***d***, bell-shaped gray areas indicate permutation distributions.

analysis testing whether it can distinguish activation patterns of the mPFC ROI during the short versus long RT blocks.

Within the mPFC ROI, the average accuracy for classifying the short and long RT blocks was 51.71%, which did not differ significantly from the theoretical chance level of 50% (Wilcoxon signed rank test, $p = 0.31$). Also, it was significantly lower than the accuracy for classifying actual self- versus other-reference blocks (average, 63.14%; paired-sample Wilcoxon signed rank test, $p = 0.002$). We additionally ran the same MVPA (short vs long RT blocks) across the whole brain with a searchlight approach, but did not find any significant area. Taken together, the difference in RTs between the self and other conditions is unlikely to explain the difference in activation patterns between the two conditions.

Which task best explains activation patterns of the self-reference condition?

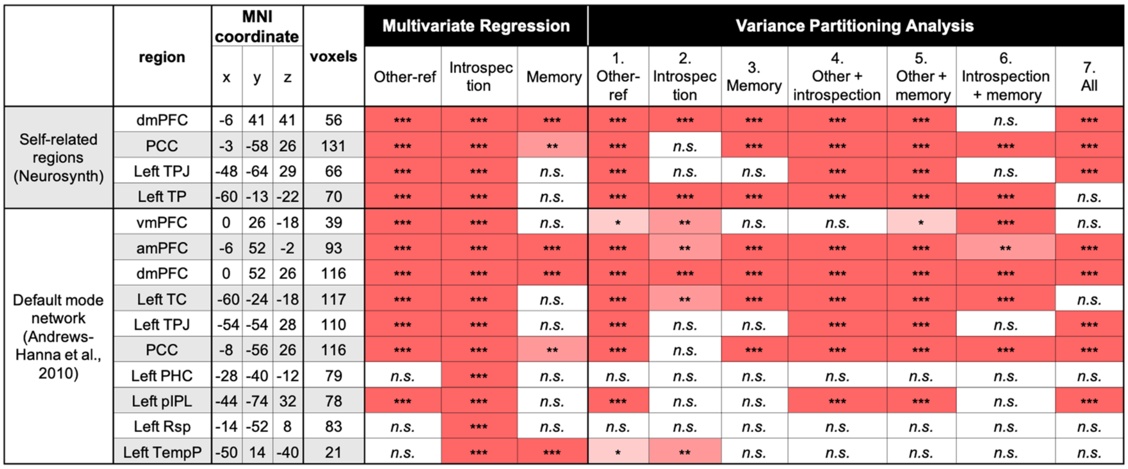
The results of the RSA reported above (Fig. 5*a–c*) indicate that mPFC activation patterns during the self-reference condition were similar to those of the other-reference, introspection, and memory conditions. However, these analyses addressed neural pattern similarity separately for each pair of tasks. To compare pattern similarity across the tasks within the same framework, we carried out a multivariate pattern regression analysis where activation patterns of the self > semantic contrast were a dependent variable, whereas those of the other > semantic, introspection > categorization, and memory > knowledge contrasts were independent variables (Fig. 3). Activation patterns of each of the three contrasts were significantly associated with mPFC ROI activation patterns of the self > semantic contrast (Fig. 6*b*; all $p_{\text{perm}} < 0.001$), suggesting that the similarity in mPFC neural responses between the self-reference task and each of the other three tasks remains significant even after controlling for the effect of neural responses during the other two tasks.

Adjusted R^2 was significantly lower than that of the noise-ceiling model ($p_{\text{perm}} < 0.001$; Fig. 6*c*). Hence, there was still unexplained variance even after considering noise in the fMRI data, suggesting that there were patterns of activations specific to the self-reference judgment (not shared by the other three tasks). Activation patterns of the other > semantic, introspection > categorization, and memory > knowledge contrasts collectively explained, on average, 79.44% of explainable variances in the mPFC activation patterns of the self > semantic contrast.

VPA

We conducted a VPA to quantify how much variance in the mPFC ROI responses of the self > semantic contrast is explained uniquely by activation patterns of each of the other > semantic, introspection > categorization, and memory > knowledge contrasts, while considered together with the other two conditions. We present the results in Figure 6*d*. Each of the seven portions significantly explained the variance in neural responses of the self > semantic contrast (all $p_{\text{perm}} < 0.001$). The results suggest that the mPFC activation patterns reflect multiple cognitive processes. For example, a significant amount of variances explained by all three contrasts indicate that there were specific patterns of mPFC neural responses that were shared across self-reference, other-reference, introspection, and memory tasks, which likely reflect a cognitive process common for the four tasks. Similarly, a significant amount of variances explained by the other-reference and introspection conditions indicate that there were specific patterns of mPFC neural responses that were shared across self-reference, other-reference, and introspection tasks, which likely reflect a cognitive process common for the three tasks, but not the memory task (see below for more discussion). These patterns of shared and unique variance among the tasks align with the RSA/MVPA results reported earlier (Fig. 5): shared variance accounts for the similarity between tasks observed in RSA (Fig. 5*a–c*), whereas unique variance explains the task-specific differences captured by MVPA (Fig. 5*d–f*).

We ran the same multivariate pattern regression and VPAs in other regions related to self-reference (based on the same Neurosynth meta-analysis map with the term "self-referential") and to the default mode network (based on Andrews-Hanna et al., 2010). The anterior and dorsal parts of the mPFC (amPFC and dmPFC) of the default mode network were the only regions that evinced the same pattern of the results as the mPFC reported above: (1) significantly positive beta values for all three independent variables (multivariate pattern regression) and (2) significantly positive variance explained for all seven



	region	MNI coordinate			voxels	Multivariate Regression			Variance Partitioning Analysis						
		x	y	z		Other-ref	Introspec tion	Memory	1. Other-ref	2. Introspec tion	3. Memory	4. Other + introspection	5. Other + memory	6. Introspection + memory	7. All
Self-related regions (Neurosynth)	dmPFC	-6	41	41	56	***	***	***	***	***	***	***	***	*n.s.*	***
	PCC	-3	-58	26	131	***	***	**	***	*n.s.*	***	***	***	***	***
	Left TPJ	-48	-64	29	66	***	***	*n.s.*	***	*n.s.*	*n.s.*	***	***	*n.s.*	***
	Left TP	-60	-13	-22	70	***	***	*n.s.*	***	***	***	***	***	***	*n.s.*
Default mode network (Andrews-Hanna et al., 2010)	vmPFC	0	26	-18	39	***	***	*n.s.*	*	**	*n.s.*	*n.s.*	*	***	*n.s.*
	amPFC	-6	52	-2	93	***	***	***	***	**	***	***	***	**	***
	dmPFC	0	52	26	116	***	***	***	***	***	***	***	***	***	***
	Left TC	-60	-24	-18	117	***	***	*n.s.*	***	**	***	***	***	***	*n.s.*
	Left TPJ	-54	-54	28	110	***	***	*n.s.*	***	*n.s.*	*n.s.*	***	***	*n.s.*	***
	PCC	-8	-56	26	116	***	***	**	***	*n.s.*	***	***	***	***	***
	Left PHC	-28	-40	-12	79	*n.s.*	***	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*
	Left pIPL	-44	-74	32	78	***	***	*n.s.*	***	*n.s.*	*n.s.*	***	***	*n.s.*	***
	Left Rsp	-14	-52	8	83	*n.s.*	***	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*
	Left TempP	-50	14	-40	21	*n.s.*	***	***	*	**	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*

Figure 7. Results of the multivariate pattern regressions and VPAs. We defined the self-related brain regions based on the Neurosynth meta-analysis map. Regions within the default mode network were based on Andrews-Hanna et al. (2010). For the self-related ROIs, we used all voxels within each cluster. For the ROIs from the default mode network, we used a 9 mm sphere surrounding the center coordinate (maximum of 123 voxels). $***p < 0.001$, $**p < 0.01$, and $*p < 0.05$ (uncorrected). All p values rely on permutation test (1,000 times). n.s. nonsignificant; dmPFC, dorsomedial prefrontal cortex; PCC, posterior cingulate cortex; TPJ, temporoparietal junction; TempP, temporal pole; vmPFC, ventromedial prefrontal cortex; amPFC, anteromedial prefrontal cortex; TC, temporal cortex; PHC, parahippocampal cortex; pIPL, posterior inferior parietal lobule; Rsp, retrosplenial cortex; TempP, temporal pole.

portions (VPA; Fig. 7), indicating a unique and complex role played by the mPFC during thinking about the self.

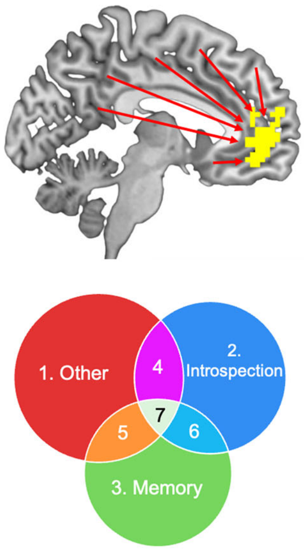
Discussion

We provided a more nuanced and precise picture of the mPFC's role during thinking about the self. Replicating prior findings, each of the self-reference, other-reference, introspection, and autographical memory tasks activated the mPFC compared with their corresponding control condition (Fig. 4). Furthermore, we demonstrated that the relationship between activation patterns during the self-reference task and those of the other three tasks (other-reference, introspection, and autobiographical memory) was intricate. That is, mPFC neural responses during the self-reference task were not simply similar to one task and different from the other two tasks. Instead, the mPFC neural responses during the self-reference task were both similar and distinct at the same time from each of the other-reference, introspection, and autobiographical memory tasks (Fig. 5). The mPFC was the only region across the whole brain that showed these patterns of results.

Furthermore, the multivariate pattern regression together with the VPAs revealed complex relationships of activation patterns of each of the three other tasks to mPFC neural responses during the self-reference task (Fig. 6). According to the VPAs, not only each of the other-reference, introspection, and memory tasks uniquely explained significant amounts of variances in mPFC neural responses during the self-reference task, but also each pair of these tasks and all three tasks jointly explained significant amounts of variances of the mPFC neural responses during the self-reference task (Fig. 6*d*). Hence, it suggests that there are cognitive processes common to thinking about the self and (1) each of the three tasks, (2) each pair of the three tasks, and (3) all three—indicating a total of at least seven overlapping cognitive components. In addition, the adjusted R^2 of the full model was significantly lower than that of the noise-ceiling model (Fig. 6*c*), suggesting that some mPFC responses during the self-reference task reflect cognitive processes that are not shared with the other tasks and may be unique to self-referential cognition. Altogether, our results indicate that there are at least eight cognitive processes (Fig. 8) at play simultaneously when performing the self-reference task, some of which are common across tasks. Thus, the present study goes beyond simply identifying similarities or differences in activation patterns. It demonstrates that the self-reference task—often treated as a unitary process—involves multiple and dissociable cognitive components. By applying VPA within the same sample of participants across multiple tasks, we were able to quantify the unique and shared contributions of each cognitive process to mPFC activity. This approach, especially when paired with MVPA and RSA, remains underutilized in this domain and represents a methodological advance for dissecting complex psychological functions.

Our VPA indicated that the mPFC is the only region that showed significantly positive variance explained for all seven portions (Figs. 6, 7). This result suggests that the mPFC, one of the core hubs of the default mode network (Andrews-Hanna et al., 2010; Andrews-Hanna, 2012), is a place where necessary information is gathered and integrated for judgments based on internally constructed representations. Yeshurun et al. (2021) considered the default mode network as an active and dynamic sense-making network that integrates incoming extrinsic information with prior intrinsic information to form rich, context-dependent models of situations as they unfold over time. More recently, Menon (2023) argued that the default mode network integrates multiple cognitive functions to create a coherent internal narrative of our experiences (see also Koban et al., 2021). Our findings offer empirical support for these frameworks by showing that the mPFC, within the default mode network, serves as a central convergence hub—integrating diverse sources of information to enable task-relevant judgments.

As a metaphor, this integration process is akin to making a soup: ingredients (information) are gathered from various parts of the kitchen (brain) and brought together in a pot (mPFC), where they are mixed and transformed into a cohesive dish (judgment). Just as many soups share core ingredients—e.g., both Italian minestrone and Japanese miso soup use common vegetables, and all soups require water—cognitive tasks often rely on shared processes, with some core mechanisms being common across many tasks. Similarly, internally directed tasks



	Cognitive process shared across self-reference and other tasks	Possible candidates
1	Cognitive process common to self-reference and other-reference tasks	• Evaluating consistency with internal goal (e.g., self-enhancement and other-enhancement) • Evaluating personal/social relevance • Mentalizing (Theory of Mind): Thinking about traits or attributes for oneself or another person • Conceptual and social information processing: Accessing and evaluating social knowledge, including how the self or another person fits within social norms
2	Cognitive process common to self-reference and introspection tasks	• Paying attention to one's internal state • Self-referential processing or self-related cognition • Internally constructing mental representations about oneself • Self-awareness and evaluation of one's own thoughts
3	Cognitive process common to self-reference and memory tasks	• Retrieval of personal memories • Conceptual and semantic integration: Accessing self-related concepts and integrating them with existing knowledge structures
4	Cognitive process common to self-reference, other-reference, and introspection tasks	• Inferential processing • Affective evaluation
5	Cognitive process common to self-reference, other-reference, and memory tasks	• Imagery and mental simulation: Visualizing past experiences or conceptualizing personal traits • Social knowledge retrieval: Accessing stored semantic and episodic knowledge about people
6	Cognitive process common to self-reference, introspection, and memory tasks	• Affective self-evaluation: Reflecting on one's own emotions and personal significance of memories or traits
7	Cognitive process common to self-reference, other-reference, introspection, and memory tasks	• Judgment based on internally constructed representations • Emotion regulation • Abstract thought and conceptual integration: Forming abstract representations about traits, experiences, or mental states rather than processing immediate sensory input
8	Cognitive process specific to self-reference	• Access to self-knowledge stored in the mPFC in terms of self-importance • Affective processing • Evaluating consistency with a self-enhancement goal

Figure 8. Schematic illustrating the proposed integrative function of the mPFC. When participants engage in the self-reference task, the mPFC integrates information from other brain regions, each of which performs distinct cognitive processes (listed here), resulting in unique activation patterns (Figs. 5–7). The VPA showed that variances explained by each of the seven portions were all significantly positive (Fig. 6*d*), suggesting that the mPFC activation pattern reflects at least seven different cognitive processes at play simultaneously (plus self-specific processes; see Discussion). The table on the right lists possible cognitive processes corresponding to each of the seven significant portions of the VPA plus self-specific processes. Note that the possible candidates we listed are purely speculative; and we do not claim that these processes are responsible for the results.

(e.g., self-reference, other-reference, introspection, and autobiographical memory) may share overlapping cognitive components, all integrated within the mPFC to support coherent mental representations. This integrative role may explain why such a wide variety of social and cognitive tasks consistently engage the mPFC. In addition to the four tasks examined here, the mPFC has been implicated in theory of mind, episodic future thinking, and even spatial navigation (Menon, 2023). These tasks all likely involve constructing internal models by drawing on multiple sources of information—further underscoring the mPFC's central role in synthesizing diverse cognitive inputs into unified, context-sensitive representations.

This view of mPFC's role in the self-reference task invites reinterpretation of prior findings. Several studies showed that mPFC activation patterns differ depending on the target person in the self-/other-reference tasks (e.g., self vs close-other vs distant other) and the dimension of person knowledge being assessed (e.g., traits, physical attributes, and social roles; Feng et al., 2018; Courtney and Meyer, 2020; Koski et al., 2020). The present study suggests that the divergent mPFC neural responses are driven by variations in the extent to which each task engages distinct types of information—and, by extension, different cognitive processes. For instance, thinking about close others and acquaintances may rely more on one's autobiographical memory, whereas thinking about unfamiliar others (e.g., celebrities) may rely more on semantic memory (Courtney and Meyer, 2020). The mPFC activation patterns are also likely to vary depending on whether a context is general or specific ("I am friendly in general vs at the university"; Martial et al., 2018) and on differences in various dimensions of distance similarity (e.g., temporal, spatial, social, and hypothetical; Tamir and Mitchell, 2011), as these judgments often rely on divergent informational sources.

Our study does not specify what cognitive processes are at play during the self-reference task (see Fig. 8 for ideas on possible candidate processes), leaving this issue open for future research. Nonetheless, as to the self-specific cognitive process, in our prior work (Levorsen et al., 2023), we reported that the self-specific activation patterns depend on the importance of the stimuli for the self-concept, and so access to this self-concept information stored in the mPFC may be responsible for the self-specific mPFC activation patterns we observed here.

Relatedly, our results highlight an important conceptual challenge for social/cognitive neuroscientists; each of many tasks used in the field involves multiple cognitive processes (or operations), and each of these processes needs to be identified to fully understand the function of the mPFC (and any other brain regions). For example, our findings indicate that the difference between the self-reference and semantic tasks is not only the level of self-referential processing, but also that there are several other additional cognitive processes involved in the self-reference task, some of which are shared with other-reference, introspection, and memory tasks (Fig. 8). Thus, rather than a traditional brain mapping approach that identifies regions activated by the self-reference > semantic contrast, a more refined approach is needed—one that links specific neural activation patterns to underlying cognitive components, rather than to entire task conditions. Although the utility of a multivariate approach over a univariate approach has been well recognized (and its methodology has been well developed; Haynes and Rees, 2006; Haxby, 2012), identifying each basic cognitive process involved in a social/cognitive task remains a challenge (see Schaafsma et al., 2015). Addressing this issue is critical for advancing our understanding of how complex mental functions are instantiated in the brain.

Finally, we note a limitation of our study. The seven tasks that we used vary in terms of visual stimuli and requirement for button responses. Granted, we made sure to match them within each task group (i.e., self, other, and semantic tasks employed the same text stimuli), and we employed contrast values with a corresponding control task (i.e., self > semantic, other > semantic, introspection > categorization, and memory > knowledge) when investigating differences in neural responses across tasks. However, it is still possible that, due to differences in stimuli and response requirement, similarities in neural responses across tasks are underestimated (Fig. 3*a–c*), whereas differences are overestimated (Fig. 3*d–f*). Although these differences are unlikely to explain the main mPFC findings (Fig. 6), future studies should use better-matched tasks to reveal roles possibly played by other brain regions.

In conclusion, the current findings enhance understanding of the mPFC and its involvement in self-referential thinking by demonstrating its unique role in integrating diverse cognitive processes. The mPFC is not merely activated by self-reference, but also shows complex activation patterns that are both similar and distinct from other cognitive tasks such as other-reference, introspection, and autobiographical memory. Taken together with the role of the mPFC within the default mode network reported previously, the findings indicate that the mPFC serves as a hub where information from various brain regions is gathered and integrated, facilitating tasks that involve constructing internal representations.

Data availability

Unthresholded group-level statistical maps are available on NeuroVault (https://neurovault.org/collections/19712/).

References

Addis DR, Wong AT, Schacter DL (2007) Remembering the past and imagining the future: common and distinct neural substrates during event construction and elaboration. Neuropsychologia 45:1363–1377.

Alink A, Walther A, Krugliak A, van den Bosch JJF, Kriegeskorte N (2015) Mind the drift—improving sensitivity to fMRI pattern information by accounting for temporal pattern drift. bioRxiv.

Anderson NH (1968) Likableness ratings of 555 personality-trait words. J Pers Soc Psychol 9:272–279.

Andrews-Hanna JR (2012) The brain's default network and its adaptive role in internal mentation. Neuroscientist 18:251–270.

Andrews-Hanna JR, Reidler JS, Sepulcre J, Poulin R, Buckner RL (2010) Functional-anatomic fractionation of the brain's default network. Neuron 65:550–562.

Buckner RL, DiNicola LM (2019) The brain's default network: updated anatomy, physiology and evolving insights. Nat Rev Neurosci 20:593–608.

Burt JB, Helmer M, Shinn M, Anticevic A, Murray JD (2020) Generative modeling of brain maps with spatial autocorrelation. Neuroimage 220: 117038.

Cai H, Wu L, Shi Y, Gu R, Sedikides C (2016) Self-enhancement among Westerners and Easterners: a cultural neuroscience approach. Soc Cogn Affect Neurosci 11:1569–1578.

Chavez RS, Heatherton TF, Wagner DD (2017) Neural population decoding reveals the intrinsic positivity of the self. Cereb Cortex 27:5222–5229.

Cooley CH (1902) *Human nature and the social order*. New York: Charles Scribner's Sons.

Courtney AL, Meyer ML (2020) Self-other representation in the social brain reflects social connection. J Neurosci 40:5616–5627.

Denny BT, Kober H, Wager TD, Ochsner KN (2012) A meta-analysis of functional neuroimaging studies of self- and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. J Cogn Neurosci 24: 1742–1752.

Feng C, Yan X, Huang W, Han S, Ma Y (2018) Neural representations of the multidimensional self in the cortical midline structures. Neuroimage 183: 291–299.

Frewen P, et al. (2020) Neuroimaging the consciousness of self: review, and conceptual–methodological framework. Neurosci Biobehav Rev 112: 164–212.

Gillihan SJ, Farah MJ (2005) Is self special? A critical review of evidence from experimental psychology and cognitive neuroscience. Psychol Bull 131: 76–97.

Goldberg II, Harel M, Malach R (2006) When the brain loses its self: prefrontal inactivation during sensorimotor processing. Neuron 50:329–339.

Gusnard DA, Akbudak E, Shulman GL, Raichle ME (2001) Medial prefrontal cortex and self-referential mental activity: relation to a default mode of brain function. Proc Natl Acad Sci U S A 98:4259–4264.

Haxby JV (2012) Multivariate pattern analysis of fMRI: the early beginnings. Neuroimage 62:852–855.

Haynes JD, Rees G (2006) Decoding mental states from brain activity in humans. Nat Rev Neurosci 7:523–534.

Heine SJ, Hamamura T (2007) In search of East Asian self-enhancement. Pers Soc Psychol Rev 11:4–27.

James W (1890) *The principles of psychology*. New York: Henry Holt and Company.

Kelley WM, Macrae CN, Wyland CL, Caglar S, Inati S, Heatherton TF (2002) Finding the self? An event-related fMRI study. J Cogn Neurosci 14:785–794.

Kim H (2012) A dual-subsystem model of the brain's default network: self-referential processing, memory retrieval processes, and autobiographical memory retrieval. Neuroimage 61:966–977.

Koban L, Gianaros PJ, Kober H, Wager TD (2021) The self in context: brain systems linking mental and physical health. Nat Rev Neurosci 22:309–322.

Koski JE, McHaney JR, Rigney AE, Beer JS (2020) Reconsidering longstanding assumptions about the role of medial prefrontal cortex (MPFC) in social evaluation. Neuroimage 214:116752.

Kriegeskorte N, Goebel R, Bandettini P (2006) Information-based functional brain mapping. Proc Natl Acad Sci U S A 103:3863–3868.

Kurdi B, Lozano S, Banaji MR (2017) Introducing the open affective standardized image set (OASIS). Behav Res Methods 49:457–470.

Legrand D, Ruby P (2009) What is self-specific? Theoretical investigation and critical review of neuroimaging results. Psychol Rev 116:252–282.

Levorsen M, Aoki R, Matsumoto K, Sedikides C, Izuma K (2023) The self-concept is represented in the medial prefrontal cortex in terms of self-importance. J Neurosci 43:3675–3686.

Levorsen M, Ito A, Suzuki S, Izuma K (2021) Testing the reinforcement learning hypothesis of social conformity. Hum Brain Mapp 42:1328–1342.

Martial C, Stawarczyk D, D'Argembeau A (2018) Neural correlates of context-independent and context-dependent self-knowledge. Brain Cogn 125:23–31.

Martinelli P, Sperduti M, Piolino P (2013) Neural substrates of the self-memory system: new insights from a meta-analysis. Hum Brain Mapp 34:1515–1529.

Menon V (2023) 20 years of the default mode network: a review and synthesis. Neuron 111:2469–2487.

Moran PA (1950) Notes on continuous stochastic phenomena. Biometrika 37: 17–23.

Moran JM, Macrae CN, Heatherton TF, Wyland CL, Kelley WM (2006) Neuroanatomical evidence for distinct cognitive and affective components of self. J Cogn Neurosci 18:1586–1594.

Murray RJ, Schaer M, Debbane 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.

Nakao T, Ohira H, Northoff G (2012) Distinction between externally vs. internally guided decision-making: operational differences, meta-analytical comparisons and their theoretical implications. Front Neurosci 6:31.

Nichols TE, Holmes AP (2002) Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum Brain Mapp 15:1–25.

Nili H, Wingfield C, Walther A, Su L, Marslen-Wilson W, Kriegeskorte N (2014) A toolbox for representational similarity analysis. PLoS Comput Biol 10:e1003553.

Northoff G (2016) Is the self a higher-order or fundamental function of the brain? The "basis model of self-specificity" and its encoding by the brain's spontaneous activity. Cogn Neurosci 7:203–222.

Parelman JM, Dore BP, Cooper N, O'Donnell MB, Chan HY, Falk EB (2022) Overlapping functional representations of self- and other-related thought are separable through multivoxel pattern classification. Cereb Cortex 32: 1131–1141.

Ritchie JB, Bracci S, Op de Beeck H (2017) Avoiding illusory effects in representational similarity analysis: what (not) to do with the diagonal. Neuroimage 148:197–200.

Schaafsma SM, Pfaff DW, Spunt RP, Adolphs R (2015) Deconstructing and reconstructing theory of mind. Trends Cogn Sci 19:65–72.

Shulman GL, Fiez JA, Corbetta M, Buckner RL, Miezin FM, Raichle ME, Petersen SE (1997) Common blood flow changes across visual tasks: II. Decreases in cerebral cortex. J Cogn Neurosci 9:648–663.

Summerfield JJ, Hassabis D, Maguire EA (2009) Cortical midline involvement in autobiographical memory. Neuroimage 44:1188–1200.

Tamir DI, Mitchell JP (2011) The default network distinguishes construals of proximal versus distal events. J Cogn Neurosci 23:2945–2955.

Thornton MA, Mitchell JP (2018) Theories of person perception predict patterns of neural activity during mentalizing. Cereb Cortex 28:3505–3520.

Wagner DD, Chavez RS, Broom TW (2019) Decoding the neural representation of self and person knowledge with multivariate pattern analysis and data-driven approaches. Wiley Interdiscip Rev Cogn Sci 10:e1482.

Wake SJ, Izuma K (2017) A common neural code for social and monetary rewards in the human striatum. Soc Cogn Affect Neurosci 12: 1558–1564.

Wen T, Mitchell DJ, Duncan J (2020) The functional convergence and heterogeneity of social, episodic, and self-referential thought in the default mode network. Cereb Cortex 30:5915–5929.

Woo CW, Koban L, Kross E, Lindquist MA, Banich MT, Ruzic L, Andrews-Hanna JR, Wager TD (2014) Separate neural representations for physical pain and social rejection. Nat Commun 5:5380.

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.

Yarkoni T, Poldrack RA, Nichols TE, Van Essen DC, Wager TD (2011) Large-scale automated synthesis of human functional neuroimaging data. Nat Methods 8:665–670.

Yeshurun Y, Nguyen M, Hasson U (2021) The default mode network: where the idiosyncratic self meets the shared social world. Nat Rev Neurosci 22: 181–192.