Overlapping Functional Representations of Self and Other Related Thought Are Separable Through Multi-Voxel Pattern Classification.

Jacob M. Parelman[1,*], Bruce P. Doré[2], Nicole Cooper[1], Matthew Brook O'Donnell[1], Hang-Yee Chan[3], and Emily B. Falk[1]

[1] Annenberg School for Communication, University of Pennsylvania; [2] Desautels Faculty of Management, McGill University; [3] Rotterdam School of Management, Erasmus University Rotterdam.

Annenberg School for Communication, University of Pennsylvania.

3620 Walnut Street

Philadelphia, PA. 19104. USA

(816) 651-9390

Jacob.parelman@asc.upenn.edu; Emily.falk@asc.upenn.edu

Running Title: Self-Other Representations Separable Through MVPA

Abstract

Self-reflection and thinking about the thoughts and behaviors of others are important skills for humans to function in the social world. These two processes overlap in terms of the component processes involved, and share overlapping functional organizations within the human brain, in particular within the medial prefrontal cortex (MPFC). Several functional models have been proposed to explain these two processes, but none has directly explored the extent to which they are distinctly represented within different parts of the brain. This study used multi-voxel pattern classification to quantify the separability of self- and other-related thought in the MPFC and expanded this question to the entire brain. Using a large-scale mega-analytic dataset, spanning three separate studies (n=142), we find that self- and other-related thought can be reliably distinguished above chance within the MPFC, posterior cingulate cortex and temporal lobes. We highlight sub-components of the ventral medial prefrontal cortex that are particularly important in representing self-related thought, and sub-components of the orbitofrontal cortex robustly involved in representing other-related thought. Our findings indicate that representations of self- and other-related thought in the human brain are described best by a distributed pattern rather than stark localization or a purely ventral to dorsal linear gradient in the MPFC.

Keywords: Classification, MPFC, MVPA, Self-reference, Social

Introduction

A hallmark of human cognition is the ability to reflect upon one's thoughts and behaviors, which inform one's own self-image (James 1890; Gallagher 2000). Similarly, humans are also adept social learners, observing others' behaviors, inferring their mental states, and using this to inform personal impressions (Frith and Frith 2006; Ma et al. 2014). Whether and how judgements about the self and others are distinct is a question that has interested psychologists for decades (Rogers et al. 1977; Bower and Gilligan 1979). Neuroimaging research has more recently contributed to this question, and identified regions of the brain associated with self-related and other-related thought. The extent to which these neural correlates overlap, however, is still not fully understood. Systematic reviews, individual studies, and meta-analytic research highlight that the medial prefrontal cortex (MPFC), in particular, is involved in both self- and other-related thought (Wagner et al. 2019; Bergstrom et al. 2015; Jenkins et al. 2008; Pfeifer et al. 2007), but that there is likely not a stark sub-localization of these processes (Denny et al. 2012). Instead of a binary distinction between these two processes, it is likely that they are represented grossly similarly in some parts of the brain, but it is not known how distinguishable they are in terms of more granular representations. Therefore, this research aims to estimate where, and to what extent, self- and other-related thought are distinctly represented in the MPFC, and throughout the brain as a whole.

Neural correlates of self-related thought

There are several regions of the brain that are associated with self-related processing, including the posterior cingulate cortex, precuneus and MPFC (Ochsner et al. 2004; Mitchell et al. 2005; Kelley et al. 2002; Jenkins and Mitchell 2010; Martial et al. 2018; Johnson et al. 2006).

Whether individuals are assessing emotional content or episodes related to the self (Ochsner et al. 2004; De Pisapia et al. 2019; Verfaellie et al. 2019), the similarity of others' faces to their own (Mitchell et al. 2005) or their own personality traits (Heatherton et al. 2006; D'Argembeau et al. 2007; Martinelli et al. 2013, Beer et al. 2010; Rameson et al. 2010), the MPFC is reliably activated.

The MPFC is a relatively large area of the brain, and both dorsal and ventral portions of the MPFC have been reported in separate research studies of self-related thought (van der Meer et al. 2010; Ochsner et al. 2004). Some research, however, argues that self-related processing is most robustly localized to ventral portions of the MPFC (VMPFC) (van der Meer et al. 2010; Wagner et al. 2012). This position is bolstered by the fact that patients with damage to the VMPFC also show significant deficits in their ability to remember self-relevant information and memories (Wheeler et al. 1997; Philippi et al. 2012). Indeed, individuals with focal damage to the VMPFC have notable difficulty in remembering information about themselves, but do not show similar difficulties in remembering information about other people (Marquine et al. 2016). The partly specialized role for the VMPFC in processing information about the self, relative to others, is further specified by patients with lesions to regions other than the VMPFC (e.g. lateral occipital, temporal and parietal regions) who do not show similar deficits in self-referential processing (Philippi et al. 2012).

Neural correlates of other related thought

The MPFC is also activated when individuals make trait judgments about other people (Mitchell et al. 2002), form impressions of others (Mende-Siedlecki et al. 2013), hold social information in working memory (Meyer and Collier 2020), and when they infer the mental states

of others (Hampton et al. 2008; Saxe 2006; Skerry and Saxe 2014; Blakemore 2008). A large body of work has shown that the dorsal medial prefrontal cortex (DMPFC) is more strongly associated with the processing of information relevant to other people than the self (see Van Overwalle 2009 for review). Further, direct contrasts of thinking about others and thinking about oneself have revealed greater activation in the DMPFC, supporting the specificity hypothesis (D'Argembeau et al. 2007). The DMPFC also exhibits increased coherence with other regions of the default mode network (e.g. posterior cingular cortex (PCC), temporoparietal junction (TPJ)) when assessing the traits of others but not oneself, further linking this sub-region of the MPFC to social cognition (Hassabis et al. 2014).

More broadly, thinking about the traits and behaviors of others and inferring their thoughts are also all -- to different degrees -- associated with activation in the temporal parietal junction, the precuneus, and superior temporal sulcus. (Saxe 2006; Dodell-Feder et al. 2011; Saxe and Waxler 2005). These regions also provide useful information in the prediction of socially relevant cognition, like assessing the social status of other individuals (Parkinson et al. 2017).

Overlap of self and other-related thought

Other neuroimaging literature, however, provides reason to question whether self- and other-related thought can be starkly linked to distinct portions of the MPFC (Saxe et al. 2006). For example, ventral regions of the MPFC are frequently identified in studies of socially oriented thought (D'Argembeau et al. 2005; Kelley et al. 2002; Lou et al. 2004; Koster-Hale et al. 2017) and dorsal regions are similarly activated during self-oriented thought (Seger et al. 2004; Craik et al. 1999; Schmitz et al. 2004; Gusnard et al. 2001). Such double associations have motivated

alternative models of organization within the MPFC. More specifically, researchers have suggested that self- and other-related thought may be organized along a ventral to dorsal gradient within the MPFC (Heatherton et al. 2006; Mitchell et al. 2005; Tamir and Mitchell 2010). Perhaps the most influential evidence for this perspective comes from a meta-analysis conducted by Denny and colleagues (2012). In this study the authors conducted a review of the literature on self-related thought and other-related thought, and found 47 peak activations for self-thought and 43 peak activations for other-thought across 107 studies within the MPFC. The authors then conducted a logistic regression using the z-coordinate of each activation to model the organization of this body of research. They found a ventral-dorsal gradient, such that peak activations located more ventrally were more likely to be (though not exclusively) reported in a study of self-related thought and activations located more ventrally were more likely to be (though not exclusively) reported in a study of other-related thought (Denny et al. 2012).

Denny et al. (2012) provide a parsimonious probabilistic model for how self- and other-related thought are distributed within the MPFC. What this model does not answer, however, is whether, and to what extent, these two processes can be distinguished within sub-regions of the MPFC. Indeed, even within the data described by Denny and colleagues, there is substantial overlap in where studies of self- and other-related thought report activations. This suggests that in addition to a dorsal-ventral gradient, there may be varying degrees of overlap in the function of specific regions. In order to directly test the question of separability within sub-regions, it is necessary to compare direct observations of self-related and other-related thought within the MPFC; a level of analysis which meta-analysis abstracts from, and which requires more direct access to raw data. Denny et al., reveals that, indeed, self- and other-related thought are

distributed throughout the MPFC, but do not directly quantify the extent to which these distributed representations are separable.

Quantifying the separability of self- and other-related thought

The introduction of multivariate pattern analysis (MVPA) and machine learning in neuroimaging research has made it possible to identify differences between cognitive states through the analysis of their distributed patterns of activation. In one of the foundational MVPA studies, Haxby and colleagues found that multivariate patterns within the ventral temporal (VT) cortex were more sensitive to differences between different object categories compared to univariate analysis (Haxby et al. 2001; Norman et al. 2006). Activation related to different object categories in the VT overlapped substantially, but the distributed patterns related to each were discernible. Analogously, self- and other-related processing overlap substantially in the MPFC, but it is not yet clear to what extent the patterns of activation related to each are distinguishable.

Recently there has been a growing interest in applying multivariate approaches to the study of self-related processing and other-related processing (Wagner et al. 2019; Courtney and Meyer 2020; Koski et al. 2020). This research has brought new understanding to the overlap of experienced and vicarious pain, and has found that the pattern of activation elicited by one is notably similar to the other (Corradi-Dell'Acqua et al. 2016; Krishnan et al. 2016). More specific to the aims of the current analysis, Oosterwijk et al (2017), found that a model trained to classify different forms of self-related processing (e.g. thinking about one's own emotions, actions or physical feelings) is also able to classify these different forms of thinking when they are about other people (Oosterwijk et al. 2017). The evidence thus far indicates that self- and other-related processing are not only overlapping in *where* they activate the cortex but also, at least partially,

in *how* they activate the cortex. In the current study, we ask not how well a model trained on different forms of self-related processing generalizes to other-related processing, but instead ask whether, and to what extent self-related processing can be distinguished from other-related processing.

Here we employ a mega-analytic approach to this question, which integrates raw data from multiple studies. Specifically, in this study we analyze the data from a large sample of participants from three separate studies ((Falk et al. 2015; Cooper et al. 2015); one unpublished study) to ask how separable self- and other-related thought are within the MPFC and how these processes are functionally organized. We employ machine learning techniques including regularization and cross-validation to quantify the extent to which these processes are distinct in the MPFC in particular as well as throughout the whole brain. We ask four directed questions regarding the functional organization of self- and other-related processing within the brain. First, we ask whether information from the MPFC is sufficient to reliably distinguish between self- and other-related thought. Second, we ask if the derived organization of regions that distinguish between self and other-related thought in the MPFC follows a linear pattern along a ventral-dorsal gradient, or whether this organization is more complex. Third, we move beyond the MPFC to explore whether including information from the entire brain significantly improves the characterization and prediction of self- and other-related thought. Finally, we explore the map of voxel weights from the whole brain to ask how self- and other-related thought are organized outside of the MPFC.

Materials and Methods

Participants. One hundred and forty-two (N = 142) participants were included in this analysis. Participant data were collected as part of three separate neuroimaging studies (N_{study1} = 60; N_{study2} = 39; N_{study3} = 43). These three neuroimaging experiments included tasks which probed the neural substrates of self- and other-related thought among other tasks unique to each study. Participants self-identified as 55% women and 45% men. Participants' ages ranged from 18 to 77 with an average age of 29 years (SD = 12.61). Participants self-identified as: 10% Asian, 15% Black, 6% Latino, 5% multiracial, 56% White, 8% other ethnicity. The experimental protocols for each study were approved by the relevant university Institutional Review Boards (University of Pennsylvania and University of Michigan). Informed written consent was obtained from all participants.

Task and Procedure. Participants were recruited for three different neuroimaging experiments, all of which included as part of their procedures a paradigm called the self-localizer task. The self-localizer task is a widely used task to localize areas of the brain associated with self-related thought (Schmitz and Johnson 2006). Participants in the self-localizer task are presented with trait adjectives and are asked to make judgments about each word after its presentation. In all three of the current experiments, participants were asked to make judgments about the relevance of words to themselves, the relevance of words to another person, or whether the words had positive or negative valence. One of the three experiments (study 3) had additional conditions that varied slightly from the other two (described below), but the judgments focused on in the current analysis (relevance to self, relevance to another person) were shared across all three.

In studies 1 (Falk et al. 2015) and 2 (unpublished), participants made trait adjective judgments under three conditions: word describes you, word describes then-president Barack Obama, and word is positive or negative. Participants completed six blocks of each condition, each containing five trait adjective trials, for a total of 30 trials per condition. For each block participants viewed both positive and negative trait adjectives in a pseudorandom order. Each block of words began with a screen that indicated to the participant what type of judgment they were being asked to make. Each word was then presented and the participant made a judgment about the word. The word and the participant's decision remained on the screen until 2.5 seconds had elapsed, after which the next word was presented. Participants responded "yes" or "no" or "positive" or "negative" depending on whether they were making relevance judgments or valence judgments. Both "yes" and "no" responses in the self and social judgment constitute thinking about oneself or a social target, as either requires participants to reflect on their traits (or the social target's traits) and confirm or deny whether the given adjective matches.

In the third study (Cooper et al. 2015), participants made trait adjective judgments under five conditions. Four conditions required participants to take either their own perspective or the perspective of a friend, and judge whether the word described themselves or their friend. In this way, the participant could either be making judgments about whether they thought a word described themselves, whether they thought a word described a friend, whether they thought their friend would think the word described themselves, or whether they thought their friend would think the word described their friend. In the current analysis, to parallel the first two studies, we only focus on conditions in which participants took their own perspective and judged whether the words described themselves or a friend. Participants also completed a valence judgment condition like those in the studies 1 and 2. All task procedures and timing were the

same as in studies 1 and 2 except that in this condition a total of 36 words were viewed and each block contained 6 trait adjectives, three positive and three negative.

Image acquisition. Data from study 1 were acquired on a 3 Tesla GE Signa MRI scanner. Functional images were acquired using a reverse spiral sequence (TR = 2,000ms, TE = 30ms, flip angle = 90°, 43 axial slices, FOV = 220mm, slice thickness = 3mm, voxel size = 3.44, 3.44, 3.0 mm). In-plane T1-weighted images (43 slices, slice thickness = 3mm, voxel size = 0.86, 0.86, 3.0 mm) and high resolution T1-weighted images (124 slices, slice thickness = 1.02, 1.02, 1.2 mm, SPGR) were also acquired for use in coregistration and normalization. Data from study 2 were acquired on a 3 Tesla Siemens Magnetom MRI scanner. Functional images for study 2 were recorded using a multiband sequence (TR = 1,500ms, TE=25ms, flip angle = 60°, 54 axial slices, FOV = 208 mm, slice thickness = 3mm, voxel size = 3.0, 3.0, 3.0 mm). High resolution T1-weighted images were also acquired (160 slices, voxel size = 0.9, 0.9, 1.0 mm) for use in coregistration and normalization. Finally, data from study 3 were acquired using a 3-Tesla GE Signa MRI scanner. Functional images were acquired using a reverse spiral sequence (TR = 2,000ms, TE = 30ms, flip angle = 90°, 43 axial slices, FOV = 220mm, slice thickness = 3mm, voxel size = 3.44, 3.44, 3.0 mm). In-plane T1-weighted images (43 slices, slice thickness = 3mm, voxel size = 0.86, 0.86, 3.0 mm) and high resolution T1-weighted images (124 slices, slice thickness = 1.02, 1.02, 1.2 mm, SPGR) were also acquired for use in coregistration and normalization.

fMRI Analysis. Data were preprocessed using Statistical Parametric Mapping (SPM8; Wellcome Department of Cognitive Neurology, Institute of Neurology, London, UK) for all

stages except for initial despiking, which was performed using the 3dDespike program implemented in the AFNI toolbox (Cox 1996). Pre-processing steps for all three datasets included slice time correction, realignment, coregistration with both T1-weighted images, segmentation, and normalization to the MNI-152 and resampling to the same voxel size (3mm).

Task block condition-specific estimates were calculated by running a voxelwise first-level general linear model for every participant, in which the entire block of a specific condition was modeled as its own separate regressor (performed with SPM12). This regression approach (Rissman et al. 2004), results in a separate whole brain map estimate for every self-judgment and other-judgment condition block. For our analysis of the MPFC, all first level block images were then masked using a binary grey matter image of the entire medial wall (3,596 voxels). Following Denny et al., the boundaries for this mask were $|x| < 25$, $y > 15$. Importantly, we included the OFC in our mask, unlike Denny and colleagues, who restricted their mask to voxels above z= -5 (Denny, et al., 2012). For our whole brain analysis, we also masked all first-level block images using a binary grey matter image (Shen et al. 2013). T-maps for the MPFC and whole brain data were flattened to a 1-dimensional array using custom code and the python neuroimaging toolbox, nilearn (Abraham et al. 2014).

MVPA Training and Testing. Participants' data were first shuffled and split into a training and test set using an 80/20 split, and thus containing 111 and 32 participants respectively. Individual participants' data were kept together for this train-test sample split, such that any given participant's data only existed in one of these two samples. For model training, we implemented a Ridge-PCR analysis method to differentiate self and other-related processing for the MPFC and whole brain respectively. This method closely follows Wager at al's (2011) LASSO-PCR

method, in which PCA is run prior to model tuning and training. The motivation of this approach is twofold: first, PCA is initially performed in order to account for the tendency of L1 and L2 regularization to ignore the natural covariance in fMRI data and select or downweigh non-contiguous voxels, and second PCA significantly reduces the size of the feature space during model training, which reduces the computational burden of performing permutation procedures. Our analysis diverged from the LASSO-PCR approach, in that we chose to implement a Ridge regression in order to maintain all features during model training (Hoerl and Kennard 1970). Similar to LASSO-PCR, Ridge-PCR returns linear feature weights, which allowed for easily interpretable brain maps.

PCA was performed with scikit-learn (version 0.22.1), which uses a LAPACK routine to compute the singular value decomposition (http://www.netlib.org/lapack/faq.html). Following the procedures of Wager et al., we retain a full-rank set of components for both MPFC and whole-brain analyses. There are 1,145 observations in the training data (significantly fewer observations than voxels in both MPFC and whole brain), which resulted in n-observations minus 1 principle components, or 1,144. In addition to the findings of Wager et al., classification analysis using a full-rank set of components has been shown to be a powerful method for discriminating various psychological processes from neuroimaging data (Koban et al. 2021; Chang et al. 2015, Krishnan et al. 2016).

Model development for the MPFC and whole brain respectively was performed using the python machine-learning toolbox, scikit-learn (Pedregosa et al. 2011). To train and tune classifiers, 5-fold randomized cross validation was used. A grid-search protocol was implemented to tune the L2 regularization penalty hyperparameter, which tested values between 0.001 and 1 at increments of 0.1. For the classifier the hyperparameter that resulted in the best

average validation accuracy value (within folds) was finally retrained on all training data and carried forward for out-of-sample testing with the independent, held-out test set.

Classification models were tested out-of-sample for their predictive accuracy on the remaining self and other block images in the test sample (N = 384 images across 32 test-set participants). Confidence intervals for out-of-sample test accuracy were derived using a bootstrap procedure, in which 1,000 bootstrap samples were drawn with replacement. Confidence intervals that do not include 50% (chance) are considered significant.

We additionally carried out all of the above procedures using alternative classification models (PLS-DA, SVC-PCA (Barker 2010; Platt 1999)) in order establish that our model results were not dependent on our specific methodological choices. These methods indeed returned very consistent results to our main analysis, and details of these procedures can be found in our supplementary materials.

Results

MPFC classification. To differentiate self- and other- related cognition within the MPFC a Ridge-PCR model was tuned and trained on data extracted from the MPFC. Block images were used for this procedure. To first assess the predictive capacity of these models, the within sample accuracy for each of five stratified cross-validation folds was calculated. Overall, the model performed well within the training sample with an average validation accuracy score across folds of 63.7% (min=57%, max=66%). The component weights for the trained model were used to predict whether a not-yet-observed image came from a condition under which a participant was engaging in self or other-related thought. The MPFC Ridge-PCR model achieved a mean out-of-sample test accuracy of 58.9% (CI = [54%, 64%]).

Whole brain classification. Analyses aimed at distinguishing self- vs. other-related processing were next repeated using data from the entire brain. A Ridge-PCR classifier was trained and tuned in the same fashion as the MPFC analysis using block-level images and was then tested out of sample for its predictive accuracy. Training prediction accuracy was again calculated for every fold of the 5-fold cross validation and averaged for the model. The whole brain model performed well in the training sample, achieving a training accuracy score of 71.8% (min=67%, max=75%). The whole brain Ridge-PCR classifier had a mean out of sample classification accuracy score of 67.7% (CI = [62%, 73%]). The test accuracy score of the whole brain model was also directionally more accurate than the model using MPFC alone however, the confidence intervals obtained through resampling overlap substantially.

Organization of MPFC model weights. To investigate whether there is a ventral to dorsal gradient in the organization of voxel weights for the classification of self- vs. other-related thought, we examined model weights from the Ridge-PCR model organized linearly from the most ventral to dorsal portion of the MPFC. These model weights represent the relative contribution and direction of each voxel in discriminating between self- and other related thought. To obtain this organization, we projected model weights back into their original 3-D voxel space by multiplying the PCA component matrix with the best performing model regression coefficients. We correlated (Pearson r) the ventral to dorsal position (i.e. z-coordinate) of every voxel in the MPFC with the model weight for each voxel. We found initial supporting evidence for a dorsal to ventral gradient in the organization of MPFC weights (r =

0.09, p=.03); however, this relationship was weak, suggesting the possibility of a more complex organization (figure 1).

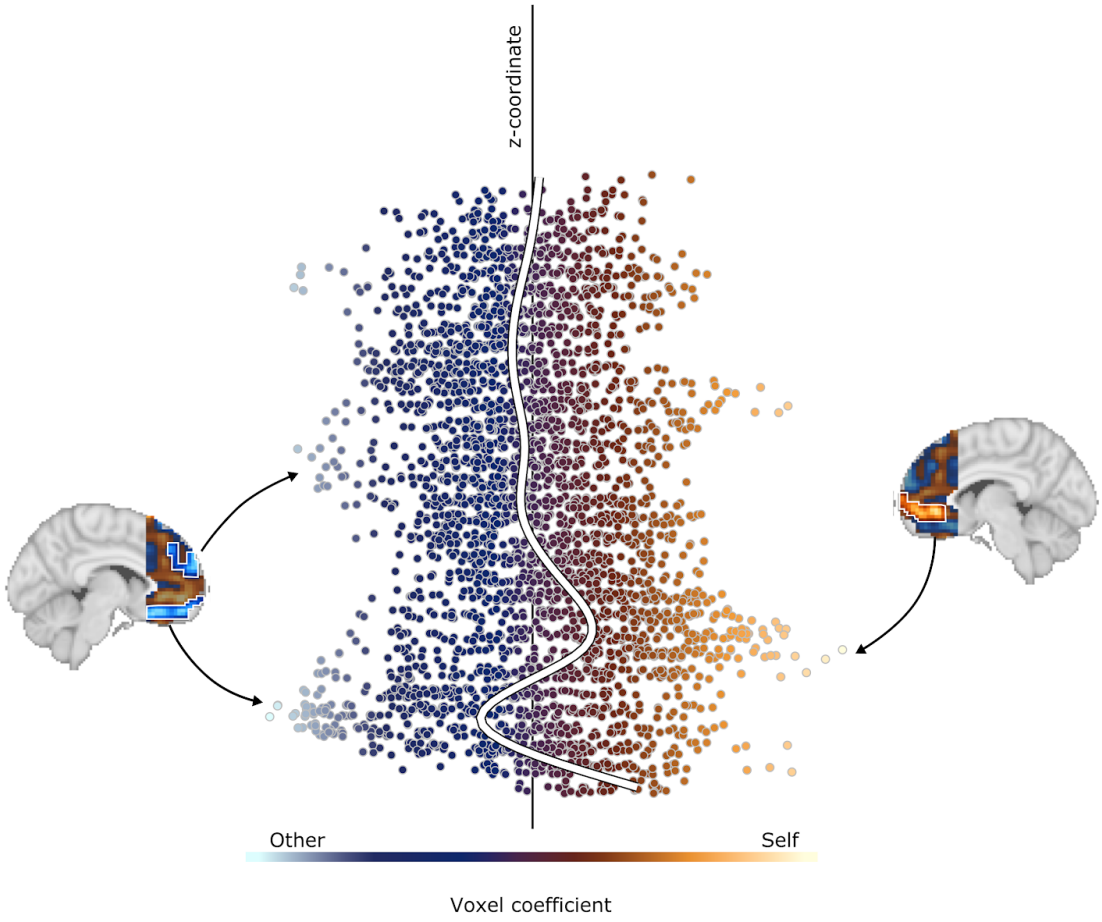

Figure 1. Organization of unthresholded voxel weights from the MPFC along the axial plane. Lighter blue indicates stronger weight and encoding for other-related processing, lighter orange indicates stronger weight and encoding for self-related processing. Z-coordinates for this graph are normalized between 0 and 1, the lowest z-coordinate in this analysis was z = -30 and the highest z = 70. Highlighted voxels indicate clusters associated with peak weights represented in the distribution. Peak weights for self-related processing were centered around z = -2, peak weights for other related processing were centered around z = -14 and z = 24. A Loess regression was used to generate the smoothed line in the figure.

We obtained a p-value to test the significance of this correlation through a permutation procedure, which involved shuffling the order of the training labels, retraining the model, projecting the trained model weights back into their original 3-D space, and computing the z-coordinate correlation 1,000 times. We also tested the reliability of the individual model weights through this same permutation procedure, resulting in a p-value for every model weight (voxel). (see supplementary materials from Pereira et al. 2009 for review of procedures). We then thresholded our model weights, controlling for multiple comparison (FDR corrected, $\alpha = 0.05$). Clusters in the OFC and VMPFC survived this thresholding procedure, as can be seen in figure 2.

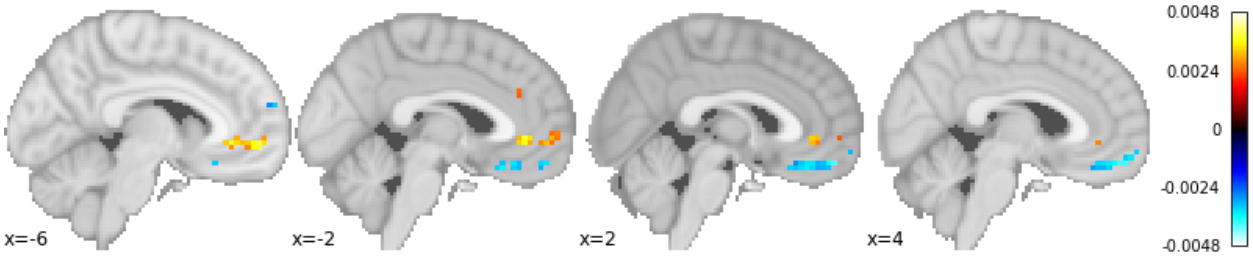

Figure 2: MPFC Ridge-PCR model weights (FDR corrected, $\alpha = 0.05$), back projected onto the whole brain. These results highlight most robust patterns differentiating self-related thought in the VMPFC, and other-related thought in the OFC and a small portion of the DMPFC. Brighter orange indicates greater weight towards encoding self-related thought, brighter blue indicates greater weight towards encoding other-related thought.

Organization of Whole Brain model weights. Voxel weights from the whole brain were also back projected into their original 3-dimensional voxel positions. Model weights, i.e. the relative contribution and direction of each voxel in discriminating between self and other related thought, were examined across the entire brain. To test the reliability of the model weights, we repeated the permutation procedures reported in the previous section for the whole brain and thresholded our model weights, controlling for multiple comparison (FDR corrected, $\alpha = 0.05$). We found that the MPFC was reliably involved in discriminating between self and other related processing as were other regions throughout the brain. Peak voxel weights coding for self-related thought were primarily localized to the VMPFC and anterior cingulate cortex (ACC). Voxels coding for other-related thought were more distributed throughout the brain, showing peak voxel weights in the PCC, left angular gyrus and left temporal lobe (figure 3).

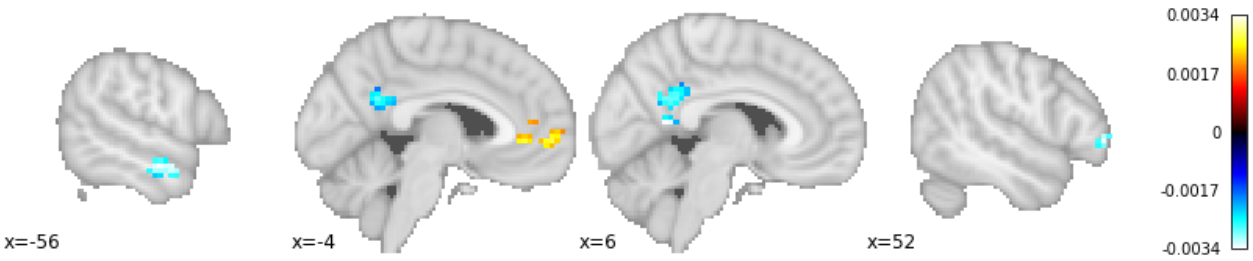

Figure 3: Whole Brain Ridge-PCR model weights (FDR corrected, $\alpha = 0.05$), back projected onto whole brain. These results highlight the most robust patterns differentiating self-related thought in the VMPFC, and other-related thought in the PCC, angular gyrus and temporal lobe. Brighter orange indicates greater weight towards encoding self-related thought, brighter blue indicates greater weight towards encoding other-related thought.

Discussion

Can patterns of activation within the brain provide reliable information for discriminating between self and other-related thought? We first focused on the MPFC, a region frequently and reliably implicated in both processes (Heatherton 2011), and then moved to examine patterns across the entire brain. Using raw data (i.e., a mega-analysis, as compared to a meta analysis; Denny et al. 2012) relevant to the question of how these processes are distinctly represented, we provide novel evidence that self- and other-related processing can be reliably distinguished using patterns of activity restricted to voxels within the MPFC. We also found that self- and other-related processing can be reliably distinguished using activity apparent across the whole brain, with differential activation in ACC and VMPFC particularly connected to self-related thought, and PCC, left angular gyrus and left temporal lobe particularly connected to other-related thought. In addition to quantitatively evaluating the explanatory power of the MPFC (and whole brain) for discriminating between self- and other-related thought, we also report data supporting the hypothesis that self- and other-related processing are functionally distributed along a ventral to dorsal gradient within the medial frontal wall (Denny et al. 2012). However, we also find that a purely linear pattern does not fully explain the effect; rather, self-related thought is most reliably represented within the middle prefrontal cortex and other-related thought within both the OFC and DMPFC, suggesting that curvilinear, rather than linear, gradient might characterize the organization of self- and other-related cognition within the medial wall of the frontal cortex.

In our analysis of the MPFC, we developed a model to discriminate between self- and other-related thought. This model performed well out-of-sample, correctly predicting 59% of our

test observations. After training and testing this model, we also investigated how the multivariate pattern of weights were organized within the MPFC, and tested whether model weights followed a linear ventral-dorsal gradient. Here, our analysis of voxel weights and their respective position along the axial plane suggests a more complicated pattern that cannot be described with a simple linear gradient. Our results highlighting a more complex pattern are also consistent with previous reviews of the MPFC that note a significant overlap in how self- and other-related processes are represented in the medial prefrontal cortex (Wagner et al. 2012), and also help explain how self- and other-related thought can appear overlapping (see Denny et al. 2012), but still be distinguishable within the MPFC.

We found that voxels that were more strongly and reliably associated with the encoding of self-related (vs. social) processing (controlling for all other activation in the MPFC) were found within the ventral portion of the MPFC and extended upwards through the ACC. This result falls in line with a wealth of previous research linking middle subregions of the MPFC with self-related thought (Heatherton 2011; Legrand and Ruby 2009). However, as shown in our whole-brain analysis (discussed below) and other research (Murray et al. 2012; Rose Addis and Tippett 2008), these areas of the MPFC are not solely responsible for encoding self-related cognition or distinguishing between thinking about oneself from thinking about others. Other regions distributed throughout the human cortex also contribute to this distinct form of cognition. These results specifically indicate that when analysis is restricted to the MPFC, a region that has shown notable functional overlap of self- and other-related thinking, middle subregions stand out as reliably encoding self-related thought.

The most influential voxels coding for other-related thought, however, were found in *both* the DMPFC and VMPFC, including the orbitofrontal cortex. Dorsal portions of the MPFC

are frequently implicated in thinking about the traits and thoughts of other people (Mitchell et al. 2005; Baron et al. 2011), and so our findings complement and extend prior understanding of this subregion. Further, supplemental analyses using the current dataset, including a voxel-wise univariate contrast and classification with average activation in ventral and dorsal portions of the MPFC, bolster our confidence in the current findings. Comparable voxel patterns to our primary findings were found in these additional analyses (supplementary materials), indicating that the organization of voxel weights found in our primary analyses are robust to analytic approaches. Lastly, we also found the organization of voxel weights to be robust when we trained the Ridge-PCR model on each of the three studies included in this mega-analysis separately. Studies 1 and 2 were most similar to the full sample model, but study 3 provided a less clear pattern; a result we attribute to its smaller sample size and alternative task design (see supplementary materials).

In addition to DMPFC, we also find that the OFC is important in reflecting on information about other people. This aligns with previous findings that the OFC is involved in social cognition (Weaverdyck et al. 2021; Beer 2006; Völlm et al. 2006), and in particular is sensitive to differences in trait judgments between the self and social targets (Hughes and Beer 2012; Beer and Hughes 2010). These findings also extend previous investigations of self- and other-related processing in MPFC (e.g. Denny et al. 2012) by further highlighting the OFC's role in social thinking. Our use of continuous data in a mega-analysis, as opposed to summary statistics focused on peak activation from previously published work, highlights a stronger role of the VMPFC in social cognition than previously emphasized.

As reflected in our findings above, and a substantial body of prior research, MPFC is centrally involved in both self- and other-related processing. However, recent work also highlights the value of incorporating whole brain patterns for characterizing psychological states

(Chang et al. 2015; Wager et al. 2020). Thus, in addition to our focus on MPFC, we also developed an additional of model, which classified self- and other-related processing using information contained within the entire brain. The inclusion of this information generated a directional improvement in the performance of our model, such that our best performing whole-brain model improved out-of-sample prediction accuracy by nearly 9% relative to our best performing MPFC model, and 18% better than chance. Previous research employing MVPA techniques to classify higher order cognitive processes, such as social perception (Brosch et al. 2013) and different aspects of self-related thought (Oosterwijk et al. 2017), tend to achieve classification accuracy scores of roughly 10-15% greater than chance.

Using data from the entire brain also allowed us to understand the relative contribution of the MPFC in encoding self- and other-related thought. Comparing the voxel weights from our best performing MPFC model and best performing whole-brain model, we saw that many of the most important voxels identified in the MPFC model remained among the most predictive within the whole-brain model. These consistent feature weights lend even stronger support for the notion that the MPFC is a critical brain region in processing both cognition relating to oneself and cognition relating to other people. The whole-brain results also provide additional insights into what other areas of the brain are important for processing information about other people relative to information about the self. Specifically, consistent with past meta-analyses (e.g., (Murray et al. 2012; Denny et al. 2012)) we found that regions associated with the default mode network (e.g. PCC, temporal lobe and angular gyrus; (Raichle 2015)) also contained the most strongly weighted voxels encoding for differentiating other-related thought from self-related thought. The PCC and MPFC are associated with both self and social cognition (Wagner et al. 2019; Brewer et al. 2013; Saxe and Powell 2006; Qin and Northoff 2011; Mahy et al. 2014); we

also provide evidence from our whole brain analysis that these regions are important for processing information about other people differentially from the self.

Machine learning techniques in neuroimaging are becoming increasingly popular within the field, and can be leveraged to generate externally valid models that can be iteratively refined and tested in additional contexts. Thoughtful interpretation of these models and their weights is necessary though, as they are not a direct measure of change in voxel activation in response to study manipulations. Coefficients from these types of models represent the relative contribution of brain regions in discriminating between conditions and the direction of their associations. The back-projected maps from the models presented here help to describe the relative influence of brain regions in coding for self against other related thought, and can be further contextualized by comparing them with more traditional contrast maps (see supplementary materials). In addition to careful interpretation of voxel weights, researchers must also be mindful of the bias-variance tradeoff implicit in modeling and prediction procedures, and how this affects meaningful inference. Here we have made efforts to implement several techniques like stratified k-fold cross-validation and out-of-sample bootstrapping to bolster confidence in the generalizability of our findings.

Looking forward, future multivariate methods like multivoxel searchlight analysis, will further help to understand the role of the MPFC in processing self- and other-related thought. More specifically, these methods can help to understand *where* in the MPFC these two processes are most reliably distinguishable, a question not answered by the analytic strategy taken here The current study utilized data from three independent studies, which all asked individuals to explicitly judge the traits of themselves and other people, affording a relatively large overall

dataset yielding a complementary high-precision method for identifying mechanisms of self and social cognition when compared to traditional coordinate-based meta-analyses.

Summary and conclusion

The current study found that self- and other-related thought are represented in partially distinct regions within the MPFC. Self- and other-related cognition were represented in both dorsal and ventral regions of the MPFC, and were weakly organized along the ventral-dorsal gradient previously proposed, with our results clarifying that clusters of VMPFC and OFC are also particularly important for distinguishing social thought. Whole brain analysis further confirmed that the MPFC was primarily responsible for representing these processes, but that the PCC and other regions of the default mode network were also involved in processing thoughts relating to other people. We provide here a key step in applying machine learning techniques to develop neural models of self- and other-related processing.

Funding This work was supported by The Michigan Center of Excellence in Cancer Communication Research/NIH Grant P50CA101451 [to Strecher V.J.]; NIH New Innovator Award 1DP2DA03515601 (to PI Falk E.B.); NIH/National Cancer Institute Grant 1R01CA180015-01 (to PI Falk E.B.); and a post-doctoral fellowship grant from NSF to Jason Coronel. We also acknowledge support from DARPA Social Sim Program (under Power of Ideas on the Internet; *POINT project,* #FA8650-17-C-7712, prime: CACI). The content is solely the responsibility of the authors and does not necessarily represent the official views of any of the funding agencies.

Acknowledgments We thank Christopher N. Cascio, Francis Tinney, Yoona Kang, Matthew Lieberman, Shelly Taylor, Lawrence An, Kenneth Resnicow, Victor Strecher for their work on the initial publication of study 1, and Holly Derry, Ian Moore, Angela Fagerlin, Thad Polk, Sonya Dal Cin, Sara Konrath, Kristin Shumaker, Nicolette Gregor, Alison Sagron, Jonathan Mitchell, and Hyun Suk Kim for their support in the development and preparation of study 1. We also thank Steven Tompson, Jean Vettel and Danielle Bassett for their work on the initial publication of study 3. We additionally thank Jason Coronel for his work on the initial research resulting in the data for study 2. We would lastly like to thank the members of Communication Neuroscience Lab for their support and feedback in the development of this project.

Positionality statement Mindful that our identities can influence our approach to science (Roberts et al, 2020) the authors wish to provide the reader with potentially relevant information about our backgrounds. With respect to race/ethnicity, one author self-identifies as Asian; five as white. With respect to gender identity, 4 self-identify as men, 2 as women.

Citation Diversity Statement Recent work in several fields of science has identified a bias in citation practices such that papers from women and other minority scholars are under-cited relative to the number of such papers in the field (Dworkin et al. 2020; Maliniak, Powers and Walter 2013; Caplar, Tacchella and Birrer 2017; Chakravartty Kuo, Grubbs and Mcllwain 2018; Thiem, Sealey, Ferrer, Trott and Kennison 2018; Dion, Sumner and Mitchell 2016; Zhou et al. 2020). Here we sought to proactively consider choosing references that reflect the diversity of the field in thought, form of contribution, gender, race, ethnicity, and other factors. First, we obtained the predicted gender of the first and last author of each reference by using databases that store the probability of a first name being carried by a woman (Dworkin et al. 2020; Zhou et al. 2020). By this measure (and excluding self-citations to the first and last authors of our current

paper), our references contain 24% woman(first)/woman(last), 21% man/woman, 20% woman/man, and 35% man/man. This method is limited in that a) names, pronouns, and social media profiles used to construct the databases may not, in every case, be indicative of gender identity and b) it cannot account for intersex, non-binary, or transgender people. We look forward to future work that could help us to better understand how to support equitable practices in science.

Supplementary Material

Supplementary material is available at Cerebral Cortex online, and code for this manuscript and all supplementary analyses can be found at: https://github.com/jmparelman/MVPA-SelfOther. Data for this will be made available upon request.

References

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., Varoquaux, G. 2014. Machine learning for neuroimaging with scikit-learn. Frontiers in Neuroinformatics, 8, 14.

Ambekar, A., Ward, C., Mohammed, J., Male, S., Skiena, S., 2009. Name-ethnicity classification from open sources. Proceedings of the 15th ACM SIGKDD international conference on Knowledge Discovery and Data Mining (pp. 49-58).

Barker, M., 2010. Partial least squares for discrimination: Statistical theory and implementation. LAP Lambert Academic Publishing.

Baron, S. G., Gobbini, M. I., Engell, A. D., Todorov, A., 2011. Amygdala and dorsomedial prefrontal cortex responses to appearance-based and behavior-based person impressions. Social Cognitive and Affective Neuroscience, 6(5), 572–581.

Beer, J.S., 2006. Orbitofrontal cortex and social regulation. Social Neuroscience: People thinking about thinking people., 41-62.

Beer, J. S., Hughes, B. L. 2010. Neural systems of social comparison and the "above-average" effect. NeuroImage, 49(3), 2671–2679.

Beer, J.S., Lombardo, M.V., Bhanji, J.P., 2010. Roles of medial prefrontal cortex and orbitofrontal cortex in self-evaluation. Cognitive Neuroscience, 22(9), 2108-2119.

Bergstrom, Z.M., Vogelsang, D.A., Benoit, R.G., Simons, J.S., 2015. Reflections of Oneself: Neurocognitive Evidence for Dissociable Forms of Self-Referential Recollection. Cerebral Cortex, 25(9), 2648-2657.

Blakemore, S. J., 2008. The social brain in adolescence. Nature Reviews Neuroscience, 9(4), 267-277.

Bower, G. H., Gilligan, S. G., 1979. Remembering information related to one's self. Journal of Research in Personality, 13(4), 420–432.

Brewer, J., Garrison, K., Whitfield-Gabrieli, S., 2013. What about the "self" is processed in the posterior cingulate cortex? Frontiers in Human Neuroscience, 7, 647.

Caplar, N., Tacchella, S., Birrer, S., 2017. Quantitative evaluation of gender bias in astronomical publications from citation counts. Nature Astronomy, 1(6), 1-5.

Chakravartty, P., Kuo, R., Grubbs, V., McIlwain, C., 2018. # CommunicationSoWhite. Journal of Communication, 68(2), 254-266.

Chang, L. J., Gianaros, P. J., Manuck, S. B., Krishnan, A., Wager, T. D., 2015. A Sensitive and Specific Neural Signature for Picture-Induced Negative Affect. PLoS Biology, 13(6), e1002180.

Cooper, N., Tompson, S., O'Donnell, M. B., Falk, E. B., 2015. Brain Activity in Self- and Value-Related Regions in Response to Online Antismoking Messages Predicts Behavior Change. Journal of Media Psychology, 27, 93–109.

Corradi-Dell'Acqua, C., Tusche, A., Vuilleumier, P., Singer, T., 2016. Cross-modal representations of first-hand and vicarious pain, disgust and fairness in insular and cingulate cortex. Nature Communications, 7, 10904.

Courtney, A. L., Meyer, M. L., 2020. Self-Other Representation in the Social Brain Reflects Social Connection. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 40(29), 5616–5627.

Cox, R.W., 1996. AFNI: software for analysis and visualization of functional magnetic

resonance neuroimages. Computational Biomedial Research, 29(3):162-173.

Craik, F. I. M., Moroz, T. M., Moscovitch, M., Stuss, D. T., Winocur, G., Tulving, E., Kapur, S., 1999. In Search of the Self: A Positron Emission Tomography Study. Psychological Science, 10(1), 26–34.

D'Argembeau, A., Collette, F., Van der Linden, M., Laureys, S., Del Fiore, G., Degueldre, C., Luxen, A., Salmon, E., 2005. Self-referential reflective activity and its relationship with rest: a PET study. NeuroImage, 25(2), 616–624.

D'Argembeau, A., Ruby, P., Collette, F., Degueldre, C., Balteau, E., Luxen, A., Maquet, P., Salmon, E., 2007. Distinct regions of the medial prefrontal cortex are associated with self-referential processing and perspective taking. Journal of Cognitive Neuroscience, 19(6), 935–944.

Denny, B. T., Kober, H., Wager, T. D., Ochsner, K. N., 2012. A meta-analysis of functional neuroimaging studies of self- and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. Journal of Cognitive Neuroscience, 24(8), 1742–1752.

De Pisapia, N., Barchiesi, G., Jovicich, J. Cattaneo, J., 2019. The role of medial prefrontal cortex in processing emotional self-referential information: a combined TMS/fMRI study. Brain Imaging and Behavior 13, 603–614.

Dion, M. L., Sumner, J. L., Mitchell, S. M., 2018. Gendered citation patterns across political science and social science methodology fields. Political Analysis, 26(3), 312-327.

Dodell-Feder, D., Koster-Hale, J., Bedny, M., Saxe, R., 2011. fMRI item analysis in a theory of mind task. NeuroImage, 55(2), 705–712.

Dworkin, J. D., Linn, K. A., Teich, E. G., Zurn, P., Shinohara, R. T., Bassett, D. S., 2020. The extent and drivers of gender imbalance in neuroscience reference lists. arXiv preprint arXiv:2001.01002.

Falk, E. B., O'Donnell, M. B., Cascio, C. N., Tinney, F., Kang, Y., Lieberman, M. D., Taylor, S. E., An, L., Resnicow, K., Strecher, V. J., 2015. Self-affirmation alters the brain's response to health messages and subsequent behavior change. Proceedings of the National Academy of Sciences, 112(7), pp. 1977–1982).

Frith, C. D., Frith, U., 2006. The neural basis of mentalizing. Neuron, 50(4), 531–534.

Gallagher, I., I. 2000. Philosophical conceptions of the self: implications for cognitive science. Trends in Cognitive Sciences, 4(1), 14–21.

Gusnard, D.A., Akbudak, E., Shulman, G.L., Raichle, M.E., 2001. Medial prefrontal cortex and self-referential mental activity: relation to a default mode of brain function. Proceedings of the National Academy of Sciences, 98(7), 4259-4264.

Hampton, A. N., Bossaerts, P., O'Doherty, J. P., 2008. Neural correlates of mentalizing-related computations during strategic interactions in humans. Proceedings of the National Academy of Sciences of the United States of America, 105(18), 6741–6746.

Hassabis, D., Spreng, R. N., Rusu, A. A., Robbins, C. A., Mar, R. A., Schacter, D. L. 2014. Imagine all the people: how the brain creates and uses personality models to predict behavior. Cerebral Cortex, 24(8), 1979–1987.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., Pietrini, P. 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science, 293(5539), 2425–2430.

Heatherton, T. F. 2011. Neuroscience of self and self-regulation. Annual Review of Psychology, 62, 363–390.

Heatherton, T. F., Wyland, C. L., Macrae, C. N., Demos, K. E., Denny, B. T., Kelley, W. M. 2006. Medial prefrontal activity differentiates self from close others. Social Cognitive and Affective Neuroscience, 1(1), 18–25.

Hoerl, A. E., Kennard, R. W. 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics: A Journal of Statistics for the Physical, Chemical, and Engineering Sciences, 12(1), 55–67.

Hughes, B. L., Beer, J. S. 2012. Orbitofrontal cortex and anterior cingulate cortex are modulated by motivated social cognition. Cerebral Cortex, 22(6), 1372–1381.

James, W. 1890. The principles of psychology, Vol II.

Jenkins, A.C., Mitchell, J.P., 2010. Medial prefrontal cortex subserves diverse forms of self-reflection. Social Neuroscience, 6(3), 211-218.

Jenkins, A.C., Macrae, C.N., Mitchell, J.P., 2008. Repetition suppression of ventromedial prefrontal activity during judgement of self and others. Proceedings of the National Academy of Sciences, 105(11), 4507-4512.

Johnson, M.K., Raye, C.L., Mitchell, K.J., Touryan, S.R., Greene, E.J., Nolen-Hoeksema, S., 2006. Dissociating medial frontal and posterior cingulate activity during self-reflection. Social Cognitive Affective Neuroscience, 1(1), 56-64.

Kelley, W. M., Macrae, C. N., Wyland, C. L., Caglar, S., Inati, S., Heatherton, T. F. ,2002. Finding the self? An event-related fMRI study. Journal of Cognitive Neuroscience, 14(5), 785–794.

Koban, L., Lee, S., Schelski, D. S., Simon, M. C., Lerman, C., Weber, B., Kable, J.W., Plassmann, H., 2021. An fMRI-based brain marker predicts individual differences in delay discounting. bioRxiv.

Koski, J. E., McHaney, J. R., Rigney, A. E., Beer, J. S., 2020. Reconsidering longstanding assumptions about the role of medial prefrontal cortex (MPFC) in social evaluation. NeuroImage, *214*, 116752.

Koster-Hale, J., Richardson, H., Velez, N., Asaba, M., Young, L., Saxe, R., 2017. Mentalizing regions represent distributed, continuous, and abstract dimensions of others' belierfs. Neuroimage, 161, 9-18.

Krishnan, A., Woo, C.-W., Chang, L. J., Ruzic, L., Gu, X., López-Solà, M., Jackson, P. L., Pujol, J., Fan, J., Wager, T. D., 2016. Somatic and vicarious pain are represented by dissociable multivariate brain patterns. eLife, 5.

Legrand, D., Ruby, P., 2009. What is self-specific? Theoretical investigation and critical review of neuroimaging results. Psychological Review, 116(1), 252-282

Lou, H. C., Luber, B., Crupain, M., Keenan, J. P., Nowak, M., Kjaer, T. W., Sackeim, H. A., Lisanby, S. H., 2004. Parietal cortex and representation of the mental Self. Proceedings of the National Academy of Sciences of the United States of America, 101(17), 6827–6832.

Ma, N., Baetens, K., Vandekerckhove, M., Kestemont, J., Fias, W., Van Overwalle, F., 2014. Traits are represented in the medial prefrontal cortex: an fMRI adaptation study. Social Cognitive and Affective Neuroscience, 9(8), 1185–1192.

Mahy, C.E.V., Moses, L.J., Pfeifer, J.H., 2014. How and where: Theory-of-mind in the brain. Developmental Cognitive Neuroscience, 9, 68-81.

Maliniak, D., Powers, R., Walter, B. F., 2013. The gender citation gap in international relations. International Organization, 67(4), 889-922.

Marquine, M. J., Grilli, M. D., Rapcsak, S. Z., Kaszniak, A. W., Ryan, L., Walther, K., Glisky, E. L., 2016. Impaired personal trait knowledge, but spared other-person trait knowledge, in an individual with bilateral damage to the medial prefrontal cortex. Neuropsychologia, 89, 245–253.

Martial, C., Stawarczyk, D., D'Argembeau, A., 2018 Neural correlates of context-independent and cortext-dependent self-knowledge. Brain and Cognition, 125, 23-31.

Martinelli, P., Sperduti, M., Piolino, P., 2013. Neural Substrates of the Self-Memory System: New Insights from a Meta-Analysis. Human Brain Mapping, 34, 1515-1529.

Mende-Siedlecki, P., Cai, Y., Todorov, A., 2013. The neural dynamics of updating person impressions. Social Cognitive and Affective Neuroscience, 8(6), 623–631.

Meyer, M.L., Collier, E., 2020. Theory of minds: managing mental state inferences in working memory is associated with the dorsomedial substystem of the default network and social integration. Social Cognitive and Affective Neuroscience, 15(1), 63-73.

Mitchell, J. P., Banaji, M. R., Macrae, C. N., 2005. General and specific contributions of the medial prefrontal cortex to knowledge about mental states. NeuroImage, 28(4), 757–762.

Mitchell, J. P., Heatherton, T. F., Macrae, C. N., 2002. Distinct neural systems subserve person and object knowledge. Proceedings of the National Academy of Sciences of the United States of America, 99(23), 15238–15243.

Murray, R. J., 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. Neuroscience and Biobehavioral Reviews, 36(3), 1043–1059.

Norman, K. A., Polyn, S. M., Detre, G. J., Haxby, J. V., 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends in Cognitive Sciences, 10(9), 424–430.

Ochsner, K. N., Knierim, K., Ludlow, D. H., Hanelin, J., Ramachandran, T., Glover, G., Mackey, S. C., 2004. Reflecting upon feelings: an fMRI study of neural systems supporting the attribution of emotion to self and other. Journal of Cognitive Neuroscience, 16(10), 1746–1772.

Oosterwijk, S., Snoek, L., Rotteveel, M., Barrett, L. F., Scholte, H. S., 2017. Shared states: using MVPA to test neural overlap between self-focused emotion imagery and other-focused emotion understanding. Social Cognitive and Affective Neuroscience, 12(7), 1025–1035.

Parkinson, C., Kleinbaum, A. M., Wheatley, T., 2017. Spontaneous Neural Encoding of Social Network Position. Nature Human Behavior, 1(5), 1-7.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., … Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research, 12, 2825–2830.

Pereira, F., Botvinick, M., 2011. Information mapping with pattern classifiers: a comparative study. NeuroImage, 56(2), 476–496.

Pfeifer, J.H., Lieberman, M.D., Dapretto, M., 2007. "I know you are but what am I?!": Neural bases of self- and social knowledge retrieval in children and adults. Journal of Cognitive Neuroscience, 19(8) 1323-1337.

Philippi, C. L., Duff, M. C., Denburg, N. L., Tranel, D., Rudrauf, D., 2012. Medial PFC damage abolishes the self-reference effect. Journal of Cognitive Neuroscience, 24(2), 475–481.

Platt, J., 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers, 10(3), 61-74.

Qin, P., Northoff, G., 2011. How is Our Self Related to Midline Regions and the Default-Mode Network? Neuroimage, 57(3), 1221-1233

Raichle, M. E., 2015. The brain's default mode network. Annual Review of Neuroscience, 38, 433–447.

Rameson, L.T., Satpute, A.B., Lieberman, M.D., 2010. The neural correlates of implicit and explicit self-relevant processing. Neuroimage, 50(2), 701-708.

Rissman, J., Gazzaley, A., D'Esposito, M., 2004. Measuring functional connectivity during distinct stages of a cognitive task. NeuroImage, 23(2), 752–763.

Rogers, T. B., Kuiper, N. A., Kirker, W. S., 1977. Self-reference and the encoding of personal information. Journal of Personality and Social Psychology, 35(9), 677–688.

Rose Addis, D., Tippett, L.J., 2008. The contributions of autobiographical memory to the content and continuity of identity a social-cognitive neuroscience approach. Self Continuity Individual and Collective Perspectives.

Saxe, R., 2006. Uniquely human social cognition. Current Opinion in Neurobiology, 16(2), 235–239.

Saxe, R., Powell, L. J., 2006. It's the thought that counts: specific brain regions for one component of theory of mind. Psychological Science, 17(8), 692–699.

Saxe, R., Moran, J. M., Scholz, J., Gabrieli, J., 2006. Overlapping and non-overlapping brain regions for theory of mind and self reflection in individual subjects. Social Cognitive and Affective Neuroscience, 1(3), 229–234.

Saxe, R., Waxler, A., 2005. Making sense of another mind: the role of the right temporo-parietal junction. Neuropsycholgia, 43(10), 1391-1399.

Schmitz, T. W., Kawahara-Baccus, T. N., Johnson, S. C., 2004. Metacognitive evaluation, self-relevance, and the right prefrontal cortex. NeuroImage, 22(2), 941–947.

Seger, C. A., Stone, M., Keenan, J. P., 2004. Cortical Activations during judgments about the self and an other person. Neuropsychologia, 42(9), 1168–1177.

Shen, X., Tokoglu, F., Papademetris, X., Constable, R. T., 2013. Groupwise whole-brain parcellation from resting-state fMRI data for network node identification. NeuroImage, 82, 403–415.

Skerry, A.E., Saxe, R., 2014. A common neural code for perceived and inferred emotion. The Journal of Neuroscience, 34(48), 15997-16008

Sood, G., Laohaprapanon, S., 2018. Predicting race and ethnicity from the sequence of characters in a name. arXiv preprint arXiv:1805.02109.

Tamir, D. I., Mitchell, J. P., 2010. Neural correlates of anchoring-and-adjustment during mentalizing. Proceedings of the National Academy of Sciences, 107(24), 10827–10832.

Thiem, Y., Sealey, K. F., Ferrer, A. E., Trott, A. M., Kennison, R., 2018. Just Ideas? The Status and Future of Publication Ethics in Philosophy: A White Paper. Technical report.

van der Meer, L., Costafreda, S., Aleman, A., David, A. S., 2010. Self-reflection and the brain: a theoretical review and meta-analysis of neuroimaging studies with implications for schizophrenia. Neuroscience and Biobehavioral Reviews, 34(6), 935–946.

Van Overwalle, F., 2009. Social cognition and the brain: a meta-analysis. Human Brain Mapping, 30(3), 829–858.

Verfaellie, M., Wank, A.A., Reid, A.G., Race, E., Keane, M.M., 2019. Self-related processing and future thinking: Distinct contributions of ventromedial prefrontal cortex and the medial temporal lobes. Cortex, 115, 159-171.

Völlm, B.A., Taylor, A.N.W., Richardson, P., Corcoran, R., Stirling, J., McKie, S., Deakin, J.F.W., Elliot, R., 2006. Neuronal correlates of theory of mind and empathy: A functional magnetic resonance imaging study in a nonverbal task. Neuroimage, 28(1), 90-98.

Wager, T. D., Atlas, L. Y., Leotti, L. A., Rilling, J. K., 2011. Predicting individual differences in placebo analgesia: contributions of brain activity during anticipation and pain experience. Journal of Neuroscience, 31(2), 439-452.

Wager, T., Van Oudenhove, L., Kragel, P., Dupont, P., Ly, H. G., Pazmany, E., Enzlin, P., Rubio, A., Delon-Martin, C., Bonaz, B., Aziz, Q., Tack, J., Fukudo, S., Kano, M., 2020. Common and distinct neural representations of aversive somatic and visceral stimulation in healthy individuals. Research Square.

Wagner, D. D., Chavez, R. S., Broom, T. W., 2019. Decoding the neural representation of self and person knowledge with multivariate pattern analysis and data-driven approaches. Wiley Interdisciplinary Reviews. Cognitive Science, 10(1), e1482.

Wagner, D. D., Haxby, J. V., Heatherton, T. F., 2012. The representation of self and person knowledge in the medial prefrontal cortex. Wiley Interdisciplinary Reviews. Cognitive Science, 3(4), 451–470.

Weaverdyck, M.E., Thornton, M.A., Tamir, D.I., 2021. The representational structure of mental states generalizes across target people and stimulus modalities. Neuroimage, 238, 118258.

Wheeler, M. A., Stuss, D. T., Tulving, E., 1997. Toward a theory of episodic memory: the frontal lobes and autonoetic consciousness. Psychological Bulletin, 121(3), 331–354.

Zhou, D., Cornblath, E. J., Stiso, J., Teich, E. G., Dworkin, J. D., Blevins, A. S., Bassett, D. S., 2020. Gender Diversity Statement and Code Note-Book v1. 0.