OXFORD

Social Cognitive and Affective Neuroscience, 2017, 1859–1868

doi: 10.1093/scan/nsx112
Advance Access Publication Date: 10 October 2017
Original article

An anterior–posterior axis within the ventromedial prefrontal cortex separates self and reward

Alla Yankouskaya,[1] Glyn Humphreys,[2] Moritz Stolte,[3] Mark Stokes,[2] Zargol Moradi,[2] and Jie Sui[4]

[1]Department of Psychology, Liverpool Hope University, Liverpool L16 9JD, UK, [2]Department of Experimental Psychology, University of Oxford, Oxford OX1 3UD, UK, [3]Department of Psychology, University of Roehampton, London SW15 5PU, UK, and [4]Department of Psychology, University of Bath, Bath BA2 7AY, UK

Correspondence should be addressed to Alla Yankouskaya, Department of Psychology, Liverpool Hope University, Hope Park, Liverpool L16 9JD, UK. E-mail: yankoua@hope.ac.uk.

Abstract

Although theoretical discourse and experimental studies on the self- and reward-biases have a long tradition, currently we have only a limited understanding of how the biases are represented in the brain and, more importantly, how they relate to each other. We used multi-voxel pattern analysis to test for common representations of self and reward in perceptual matching in healthy human subjects. Voxels across an anterior–posterior axis in ventromedial prefrontal cortex (vmPFC) distinguished (i) self–others and (ii) high–low reward, but cross-generalization between these dimensions decreased from anterior to posterior vmPFC. The vmPFC is characterized by a shift from a common currency for value to independent, distributed representations of self and reward across an anterior–posterior axis. This shift reflected changes in functional connectivity between the posterior part of the vmPFC and the frontal pole when processing self-associated stimuli, and the middle frontal gyrus when processing stimuli associated with high reward. The changes in functional connectivity were correlated with behavioral biases, respectively, to the self and reward. The distinct representations of self and reward in the posterior vmPFC are associated with self- and reward-biases in behavior.

Key words: MVPA; self; reward; vmPFC

Introduction

Understanding the nature of self-representations has been a core issue since the inception of experimental psychology, but we are still far from developing a full account. Recent work has made progress by evaluating self-biases in performance, which are often large and stable, and which can provide information about self-representation by showing what aspects of the self determine the biases. Specifically, self-bias effects have been established in memory (Fossati *et al.*, 2004), trait judgments (Kelley *et al.*, 2002; Denny *et al.*, 2012), face recognition (Sui *et al.*, 2006; Ma and Han, 2010) and even in simple perceptual matching (Sui *et al.*, 2012, 2014, 2015). A key issue, unresolved to this day, is whether such biases reflect the special status of the self for distinguishing each of us from other entities, or do self-biases stem from more basic drivers of behavior, such as the reward value linked to stimuli?

One influential account links self-bias effects to the underlying effects of reward: people show self-biases because self-related information is inherently rewarding and reward-based reinforcement enhances perception and attention as found for other stimuli linked to high reward (Tamir and Mitchell, 2012). Results of neuroimaging studies support this account by reporting substantial overlap between self-referential in the ventromedial prefrontal cortex (vmPFC) (see the review by Northoff, 2015). Interestingly, the overlapping activations encompass mainly the anterior–posterior direction across the vmPFC

Received: 3 January 2017; **Revised:** 10 September 2017; **Accepted:** 2 October 2017

spanning the anterior [including Brodmann's areas (BAs) 10, 11] and the posterior (BAs 32, 25) portions of the vmPFC in studies of reward (Knutson *et al.*, 2005; Kable and Glimcher, 2007; Padoa-Schioppa and Assad, 2008; Rushworth *et al.*, 2011; Smith *et al.*, 2014) and self-relevance (Kelley *et al.*, 2002; Moran *et al.*, 2006; Han and Northoff, 2008; van Buuren *et al.*, 2010). Furthermore, meta-analyses of imaging studies on self-referential effects (Northoff *et al.*, 2006; Denny *et al.*, 2012) and reward relevance (Clithero and Rangel, 2014) indicate the possibility of functional separation between self and reward in the vmPFC by showing that the anterior part of the vmPFC mediates monetary reward representations and the posterior part of the vmPFC is engaged in self-referential processing.

The findings that self-referential and reward processing show strong representations in the vmPFC opened a continuing debate about their relationship. The critical points of the debate are summarized in a seminal paper by Northoff and Hayes (2011) where the authors reviewed human and animal studies and proposed three possible models of the relationship between self and reward: (i) the integration model assuming overlap between self and reward, (ii) the segregation model where value assignment and self-specificity assignment are regarded as different processes that are regionally and temporally segregated and (iii) the parallel processing model posits that different aspects of self-specific processing may occur in parallel with aspects of reward-related processing at some levels, but assumes a complex relationship between self and reward with multiple interactions across the continuum.

To date, none of the proposed models received strong empirical support mainly because (i) cross-study comparisons are obscured by various factors such as differences in methods used for processing and analyzing the data, individual differences in functional brain anatomy; and (ii) different cognitive tasks and procedures are used in a single study to trigger self- and reward-biases (e.g. a personal relevance evaluation task and a gambling task) that limited direct comparisons between them.

Here, we attempt to overcome such limitations by using recently developed associative matching procedures (Frings and Wentura, 2014; Sui *et al.*, 2015; Sun *et al.*, 2016; Macrae *et al.*, 2017) that generate similar behavioral biases for self (*vs* others) and high monetary reward (*vs* low monetary reward) (Sui *et al.*, 2012). A unique aspect of this design is that it triggers common cognitive processes underlying a mental synthesis (Gallagher, 2000; Christoff *et al.*, 2011) of a neutral object and a person (or reward value) in time and enables direct comparisons between the effects of self- and reward relevance. Furthermore, in contrast to commonly used trait-judgment self-evaluations which reflect a need to evaluate external cues against internal representations of self in memory, the associative matching procedure does not require to shape and refine our conceptualizations of self, and, thus, eliminates response biases due to social desirability (Konstabel *et al.*, 2006), item popularity (Bäckström and Björklund, 2013) and affective meaning (Roy *et al.*, 2012).

The aim of this study is to explore the relationship between processing of self and reward associations across the anterior–posterior axis in the vmPFC by comparing activity patterns associated with self- and reward-biases. We hypothesize that the activation patterns for self and reward share significant similarity along the anterior–posterior axis in the vmPFC. Accurate cross-generalization between the activation patterns will support this hypothesis and provide evidence for the integration model (Northoff and Hayes, 2011). Alternatively, failing to cross-generalize between activation patterns for self and reward will indicate functional dissimilarity between them along the anterior–posterior axis providing support for the segregation model (Northoff and Hayes, 2011). There is also a possibility that self- and monetary reward-biases may generate distinct activation patterns in the posterior part of the vmPFC (see, for example, meta-analyses by Denny *et al.*, 2012 and Clithero and Rangel, 2014), but show greater similarity in the anterior part of the vmPFC, a region which instantiates many social cognitive processes (Burgess *et al.*, 2007).

Previous studies have shown that the vmPFC has different functional connectivity to the rest of the brain for self (e.g. Sui *et al.*, 2013) and reward (e.g. Smith *et al.*, 2014). A *post hoc* functional connectivity analysis was performed to test our assumption that the relationship between self- and reward-biases along the anterior–posterior axis in the vmPFC may be explained by differences in 'neural communications' between the vmPFC and the rest of the brain.

Materials and methods

Participants

Sixteen participants (eight males) aged between 22 and 34 were recruited for this study. The subjects reported no neurological conditions and had normal or corrected-to-normal vision. This experiment was approved by the Central University of Oxford Research Ethics Committee. All participants provided informed consent.

Task and stimuli

In this study, we used a recently developed procedure where a neutral geometric shape is 'tagged' with self/reward relevance by having people associate the shape with a social label (e.g. your name, your friend's name or high/low reward value). This procedure allows us to study how basic perceptual processing changes for a shape associated with self (or high reward) compared to shapes associated with other (or low reward) and measure the responses in a highly controlled way (Sui *et al.*, 2012).

Participants performed two shape–label matching tasks—based on personal relevance and on reward. In the self-task subjects were asked to imagine associations between geometric shapes and themselves and a friend (e.g. circle–you, square–friend). In the reward task, they were asked to make associations between shapes and reward values (e.g. hexagon–£16, triangle–£1). Four geometric shapes (circle, hexagon, square and triangle) were randomly assigned across participants to two associations in each task. In each task, participants were required to make a judgment of whether the display contained associated (matched) or re-paired (mismatched, e.g. circle–friend, square–you) shape–label combination (Figure 1B) by pressing response buttons ('match' or 'mismatch').

Prior the scanning session, participants performed a short practice block (12 trials) for each task with feedback on accuracy performance. Immediately after the practice, they performed the matching tasks in a brain scanner and used response buttons on a magnetic resonance imaging compatible response box.

The stimulus display contained a fixation cross ($0.8^\circ \times 0.8^\circ$) at the center of the screen with a shape ($3.8^\circ \times 3.8^\circ$) and label on either side of fixation. The distance between shape and label was 10°. Presentations of the shapes and labels were counterbalanced across trials. Each trial started with a fixation cross for 200 ms, followed by the stimulus display for 100 ms and a blank interval which remained for 1000 ms. Trials were separated by a

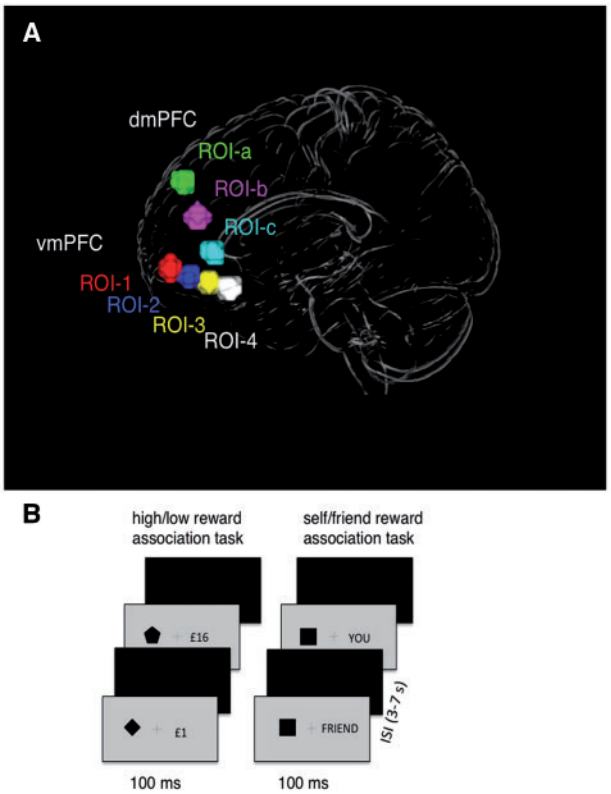

Fig. 1. (A) The ROIs defined across an anterior–posterior axis in the vmPFC and dmPFC (see coordinates in Supplementary Table S1). Within the vmPFC ROIs were selected based on prior findings showing effects of a common reward currency (ROI-1) and self-referential processing (ROI-4) (Denny *et al.*, 2012). Within the dmPFC ROIs were defined between ROI-1a (selected from studies showing sensitivity to evaluative judgments) (Mitchell *et al.*, 2006) and ROI-4. (B) The procedures for self and reward perceptual matching tasks.

jittered interstimulus interval (ranging between 2000 and 6000 ms). There were four runs of 48 trials of each task. The order of the tasks (SRSRSRSR or RSRSRSRS) was balanced across participants. For each correct answer in the reward task, participants received a reward of 2% of the amount of money displayed on a given trial. Presentation software (http://www.neurobs.com) was used to present and control the stimuli and collect behavioral measures.

ROI selection

Our primary interest focused on regions in the vmPFC, where prior studies have established effects of both reward (Schultz *et al.*, 2000; Clithero and Rangel, 2014; Smith *et al.*, 2014) and self-representation (Kelley *et al.*, 2002; Han and Northoff, 2008; Denny *et al.*, 2012; Sui *et al.*, 2013). Based on previous results we defined four regions of interest (ROIs) between the anterior and posterior parts of the vmPFC and tested the similarity between multivariate patterns for self- and reward-biases along the anterior–posterior axis bridging these ROIs (Figure 1A).

Two ROIs were defined based on the results from recent meta-analyses—one that showed 'monetary value' responses (Clithero and Rangel, 2014) (ROI-1, the most anterior part of the vmPFC) and a subcallosal region of vmPFC where self-referential effects (Denny *et al.*, 2012) have been observed (ROI-4, the most posterior part of the vmPFC) (Figure 1A). The two other ROIs (ROI-2 and ROI-3) were equally spaced and centered on the straight line between ROI-1 and ROI-4 to explore functions of the subregions of the vmPFC (the coordinates are reported in the Supplementary Table S1). All ROIs were created as spheres with a radius of 7 mm (corresponding to 57 voxels) with 2 mm gap between ROIs along the Y-axis.

To control for the results in the vmPFC we also selected an ROI in the dorsomedial prefrontal cortex (dmPFC) that is involved in evaluative prediction to reward error (Schultz *et al.*, 2000), personal evaluation (Nicolle *et al.*, 2012) and social evaluative judgments (Mitchel *et al.*, 2006). Recent research proposes that personal and reward stimuli are evaluated in the dmPFC in a similar manner as in the vmPFC (Behrens, 2013), which raises a question about whether the mechanisms at play in the dmPFC might parallel those in the vmPFC. Having functionally similar ROIs along the dmPFC as a control here provides a unique opportunity to examine (i) whether the relationship between self and reward is specific to the vmPFC and (ii) whether evaluation of the biases elicited by a common procedure yields similar effects in the dmPFC.

Similar to the vmPFC, we drew equally spaced ROIs (7 mm) placed among a superior–inferior axis between the dmPFC and the posterior ROI-4 in the vmPFC (Figure 1A, ROI-1a, ROI-2a, ROI-3a) (Supplementary Table S1).

Functional magnetic resonance imaging acquisition

Functional magnetic resonance imaging (fMRI) data were acquired on a 3T scanner (Trio, Siemens) using a 24-channel head coil. Functional images were acquired with a gradient echo T2*-weighted echo-planar sequence (TR 2000 ms, TE 30 ms, flip angle 70, 64 × 64 matrix, field of view 19.2^2 mm, voxel size 3 × 3 × 3 mm). A total of 36 axial slices (3 mm thick, no gap) were sampled for whole-brain coverage. Imaging data were acquired in eight separate 120-volume runs of 4 min 02 s each. A high-resolution T1-weighted anatomical scan of the whole brain was acquired (256 × 256 matrix, voxel size 1 × 1 × 1 mm).

fMRI pre-processing

Analysis of the imaging data was performed using SPM12 (www.fil.ion.ucl.ac.uk/spm). Functional images were realigned, unwarped, slice-timing corrected, co-registered to the participant's T1 scan, normalized to group template and smoothed with a 6 mm FWHM kernel. To reduce the intersubject anatomical variability, the group template was created based on the gray matter segmentation using DARTEL (Ashburner, 2007). The group templates were then normalized to the Montreal Neurological Institute (MNI) space and applied to each individual gray matter segmentation (see details in Supplementary Material).

The data for multi-voxel pattern analysis (MVPA) were pre-processed similar to data for ROI analysis, but without smoothing procedure to preserve the participant-specific high spatial frequency information used to index differential population codes (e.g. Haxby *et al.*, 2001; Stokes, 2015).

fMRI analyses

Three analyses were carried out: (i) ROI analysis where we examined the effects of task (self, reward) and stimulus salience [high salience (self and high reward), low salience (friend and low reward)] on the magnitude of neural responses across the ROIs (ROI-1, ROI-2, ROI-3, ROI-4, Figure 1A); (ii) multivariate pattern analysis aiming to examine whether the regions in the vmPFC shared representations for self and reward associations, and whether the neural response reliably predicts the stimuli

Fig. 2. Mean correct RTs in the self and reward perceptual matching tasks. Error bars represent ±1 SEM.

associated with self or high reward; and (iii) psychophysiological interaction (PPI) analysis to examine effects from the two biases on the relationship between ROIs in the vmPFC and other areas in the brain. In addition, to explore the whole-brain responses to self and reward stimuli, we performed a whole-brain voxel-wise analysis (the results are available in Supplementary Material).

ROI analysis. Individual fMRI time series for the self and reward runs were regressed onto a single fixed-effect general linear model to obtain parameter estimates (beta values) for each voxel across all conditions (see details in Supplementary Material, fMRI data modelling for ROI analysis). To examine the magnitude of the neural responses for learned associations we first extracted beta values for each condition in each ROI in the vmPFC (Figure 1A) for each participant from the subjects' first-level beta-maps (see ROI analysis for details) and averaged them. To test the relations between the magnitude of neural responses in the self and reward tasks, an analysis of variance (ANOVA) was conducted on the Beta values with 2 (task: self, reward) × 2 [salience: high (self, high reward), low (friend, low reward)] × 4 factors (ROI: ROI-1, ROI-2, ROI-3, ROI-4).

Multivariate pattern analysis. We used a multivariate pattern analysis based on a 'correlation approach' (Haxby, 2012; Kriegeskorte *et al.*, 2008; Nelissen *et al.*, 2013; Spaak *et al.*, 2017) where correlations between patterns of neural response serve as indices of similarity (Haxby *et al.*, 2001; Nelissen *et al.*, 2013). This approach is a variant of correlation-based nearest-neighbor classification (Williams *et al.*, 2007; Stokes *et al.*, 2009; Haxby *et al.*, 2014). The distance between two vectors in high-dimensional representational spaces reflects the cosine of the angle between the mean-centered vectors. The vectors of neural response were created using beta estimates of the entire fMRI session for each subject yielding one summary index of multivariate pattern strength per subject across the entire task.

The main advantage of using this procedure for our data is that correlations are scale- and mean-level invariant and not directly influenced by homogeneous differences in condition-specific activation. Therefore, the procedure can capture the patterns of differential neural activity, rather than magnitude differences, that constitutes the neural signature of differential population coding (e.g. Stokes *et al.*, 2009; Coutanche, 2013; Stokes, 2015).

Here we used a leave-one-run-out scheme to split the data. To perform the pattern classification, we prepared training (three of four scanning runs) and test data (the remaining scanning run) sets for each participant (see details in Supplementary Material, MVPA analysis). The correlations between the training and test sets were calculated across voxels for each ROI for each participant. A correlation that was above zero was considered as a correct classification (assigned index 1), and a correlation that was equal to or below zero was considered an incorrect classification (assigned index 0) (Stokes *et al.*, 2009). The classifier performance was evaluated using randomization tests (Ojala and Garriga, 2010).

Classification accuracy was calculated as the percentage of correct classifications across the four training–test permutations for each ROI per participant. To test whether the classification accuracy was significantly above chance level (>50%), two-tailed one sample t-tests were applied to the group data.

PPI analysis. We further used PPI analysis (O'Reilly *et al.*, 2012; Whitfield-Gabrieli and Nieto-Castanon, 2012) to examine whether self and reward processing changes functional connectivity between the vmPFC and other brain regions, in particular, we tested whether and how the four anatomical regions in the vmPFC (Figure 1A) change their connectivity with the rest of the brain in the context of two psychological factors of interest: self > high reward conditions and high reward > self-conditions (see details in Supplementary Material).

Results

Behavioral results

Accuracy for the self-task (94.25%) and for the reward task (96.95%) did not differ [$t(15) = 0.33$] and did not vary for self-values (95.6% for self *vs* 92.9% for friend) or reward values (97.8% for high reward *vs* 96.1% for low reward). Previous studies using the same task reported RT advantages for high salience stimuli (i.e. self and high reward value associations) compared to low salience stimuli (i.e. other and low reward value associations) (Sui *et al.*, 2015). The results in this study confirmed the previous findings (Figure 2). A two-factor ANOVA with task (self, reward) and value (high value, low value) as within-subject factors showed a reliable main effect of value, with RTs for high value stimuli (self, high reward) faster than for low value stimuli (friend, low reward) [$F(1, 15) = 13.7$, $P = 0.003$, $\eta_p^2 = 0.49$, 90% confidence interval (CI) [0.37, 0.63]]. There was no difference between tasks (self *vs* reward) [$F(1, 15) = 0.1$]. There was an interaction of task*value [$F(1, 15) = 5.01$, $P = 0.02$, $\eta_p^2 = 0.24$, 90% CI [0.09, 0.35]]. The RT bias to high value stimuli was greater in the self-task than the reward task although this contrast was not reliable [$t(15) = 1.67$, $P = 0.09$].

Regions of interest

Condition-specific mean beta values in ROIs are reported in the Supplementary Material. An ANOVA was performed with 2 (task: self, reward) × 2 [salience: high (self and high reward), low (friend and low reward)] × 4 (ROI: ROI-1, ROI-2, ROI-3, ROI-4) factors. There was a reliable interaction of task*ROI [$F(3, 45) = 6.81$, $P = 0.001$, $\eta_p^2 = 0.31$, 90% CI [0.17, 0.41]]. Mauchly's test for these data indicated that the assumption of sphericity had not been violated ($\chi^2 = 2.13$, $P = 0.83$). No other terms approached significance. To further test this interaction, two separate ANOVAs were performed along with polynomial contrasts on the effects of task across the four ROIs. These showed a main effect of ROI for the self-task [$F(3, 45) = 6.03$, $P = 0.005$, $\eta_p^2 = 0.29$, 90% CI [0.19, 0.43]] but not for the reward task [($F(3, 45) = 2.39$, $P = 0.08$]. Interestingly, the polynomial contrasts showed significant linear trends for both self [$F(1, 15) = 8.42$, $P < 0.05$, $\eta_p^2 = 0.36$, 90% CI [0.15, 0.46]] and reward [$F(1, 15) = 7.26$, $P < 0.05$, $\eta_p^2 = 0.33$, 90% CI [0.11, 0.42]], however the direction of the trends was different (Figure 3). Activity for the reward task decreased from the

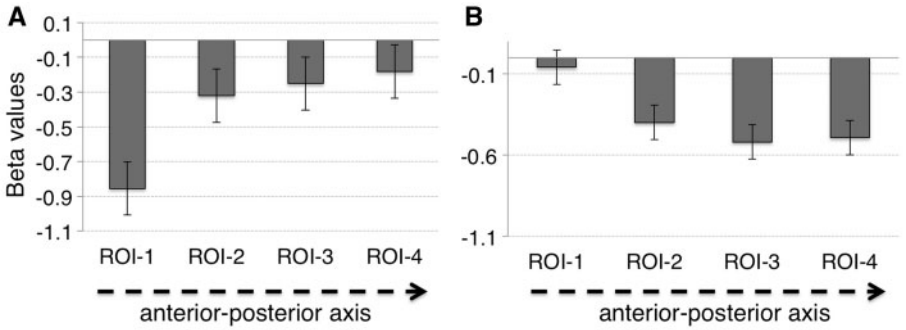

Fig. 3. Linear trend in the magnitudes of neural responses for the self-task (average self + friend) (A), and for the reward task (average high + low reward) (B).

anterior to the posterior ROIs. Activity for the self-other task increased from the anterior to the posterior ROIs. This ROI analysis indicates that there was differential engagement of the vmPFC in self and reward processing; in particular, the posterior vmPFC (pvmPFC) was more engaged in the self-task, while the anterior vmPFC (avmPFC) was more involved in the reward task.

Classification of self and high reward in the vmPFC

Within-task classification analyses revealed robust discrimination between self and friend, and between high and low reward across all ROIs. The level of classification did not differ across the tasks and ROIs [main effect of ROI, $F(3, 45) = 0.71$; main effect of task, $F(1, 15) < 1$] (Figure 4A).

Generalization of self and high reward classifications in the vmPFC

We evaluated the overlap in the neural representations for self and reward by assessing if training on one task (e.g. self *vs* friend data set) predicted successful classification of the other task (e.g. high *vs* low reward) (Figure 4B). There was a significant effect of ROI on classification accuracy [$F(3, 45) = 7.51$, $P < 0.001$, $\eta_p^2 = 0.36$, 90% CI [0.16, 0.49]] with a strong linear trend in decreasing accuracy for generalized classification along the anterior–posterior axis (from reward-ROI to self-ROI) [$F(1, 15) = 19.11$, $P = 0.001$, $\eta_p^2 = 0.57$, 90% CI [0.23, 0.68]]. There was above chance generalization of classification accuracy in the two anterior ROIs (ROI-1, ROI-2) [$t(15) = 6.01$, $P < 0.001$, $dz = 1.50$, 95% CI for dz [1.39, 1.67]; $t(15) = 2.83$, $P < 0.05$, $dz = 0.71$, 95% CI for dz [0.64, 0.88] for ROI-1 and ROI-2, respectively]; but not in the more posterior ROIs [ROI-3 and the ROI-4; $t(15) = 0.91$ and $t(15) = 0.49$, respectively].

Classification performance within the two tasks (self–self and reward–reward) was compared with that across the two tasks (self–reward and reward–self) in a 2 × 4 × 2 repeated measures ANOVA with the factors being classification (within-task *vs* across-task), ROI (ROI-1 to ROI-4) and task at test (self, reward). There were main effects of classification [$F(1, 15) = 17.51$, $P = 0.001$, $\eta_p^2 = 0.56$, 90% CI [0.26, 0.69]] and ROI [$F(3, 45) = 3.36$, $P = 0.035$, $\eta_p^2 = 0.19$, 90% CI [0.05, 0.32]], and an interaction between classification and ROI [$F(3, 45) = 4.26$, $P = 0.02$, $\eta_p^2 = 0.21$, 90% CI [0.02, 0.34]]. For within-task classification there was no 40 main effect of ROI [$F(3, 45) = 0.85$. For across-task classification there was a main effect of ROI [$F(3, 45) = 7.69$, $P < 0.001$, $\eta_p^2 = 0.38$, 90% CI [0.14, 0.45]]. Here, classification accuracy was significantly higher for ROI-1 compared to ROI-3 and ROI-4 ($P = 0.010$ and $P = 0.002$, respectively; Figure 4) after Bonferroni adjustments for multiple comparisons. There were no other significant terms (all Ps > 0.05).

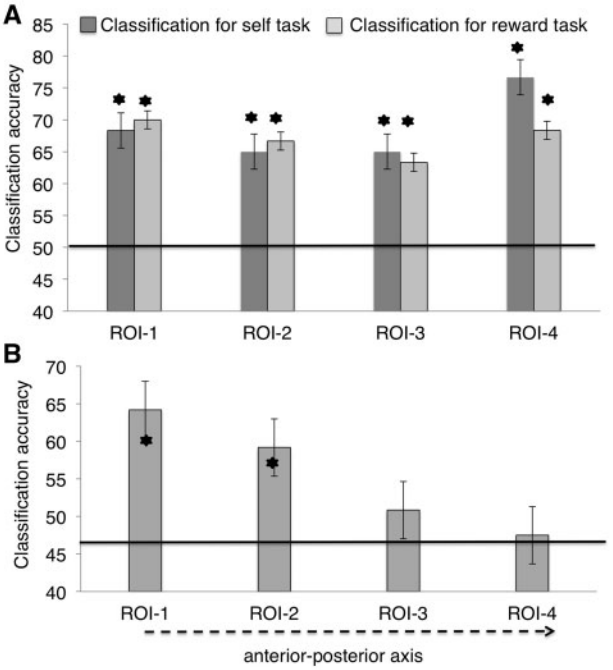

Fig. 4. Results of MVPA in the vmPFC. (A) The accuracy of classifying self *vs* other and high *vs* low reward stimuli, using leave-one-out training on each classification in the vmPFC. Error bars represent ±1 SEM. Stars denote significance at $P < 0.05$. (B) Cross-generalization classification accuracy. The classifier was trained on the reward task (or self-task) and tested on the self-task (or reward task). The results did not differ for classifications from self to reward or reward to self.

Classification accuracy and generalization of self and reward in the dmPFC

MVPA in the dmPFC showed that neither the self nor high reward could be reliably classified (Figure 5A) and there was no evidence for a shift from a common to a specific currency of value moving from the more superior to the more inferior ROIs (Figure 5B). Hence, our results are specific to the anterior–posterior axis in the vmPFC.

Functional connectivity differences between self and high reward

To test whether there was a dissociation between self and reward in terms of functional connectivity between the vmPFC and the rest of the brain, we examined effects of the high salience conditions on the relationship between the ROIs in the

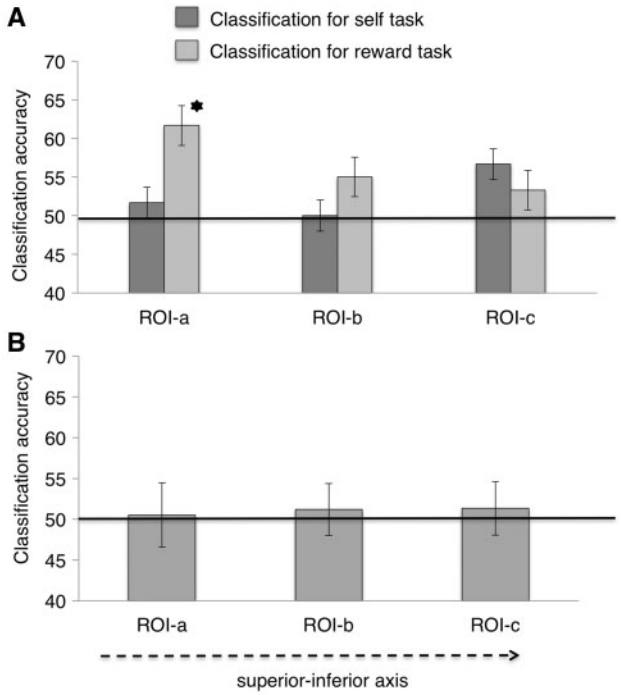

Fig. 5. Results of MVPA in the dmPFC. (A) The accuracy of classifying self *vs* other and high *vs* low reward stimuli, using leave-one-out training on each classification in the dmPFC. Error bars represent ±1 SEM. Stars denote significance at $P < 0.05$. (B) Cross-generalization accuracy classification in the dmPFC. The classifier was trained on the reward task (or self-task) and tested on the self-task (or reward task).

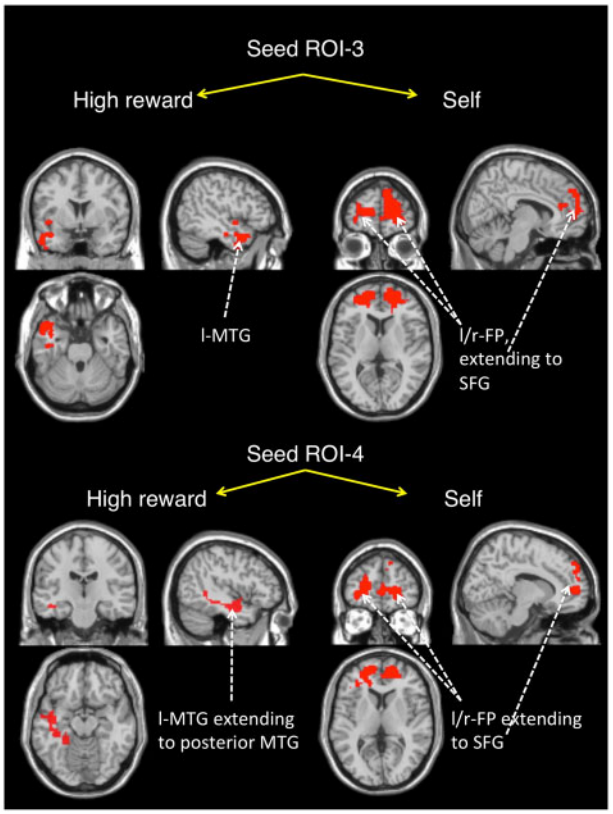

Fig. 6. PPI effects seeded in ROI-3 and ROI-4 for high reward [defined by the contrast (high reward–self)] and self [defined by the contrast (self–high reward)]. A mask of the clusters showing a PPI effect (*t*-test, a cluster corrected FDR-threshold of $P < 0.05$) is overlaid on an MNI single-subject T1. The effect size for each cluster was calculated using Cohen's *d* (Supplementary Table S2).

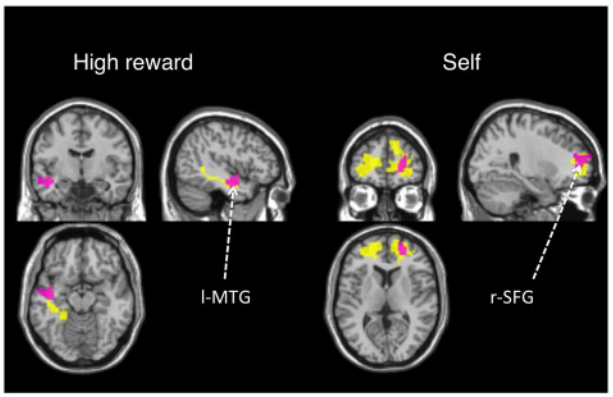

Fig. 7. Correlated behavioral RT biases and PPI results are shown in purple. For reference, these results are overlaid on PPI effects for high reward and self (in yellow from Figure 6). A mask of clusters with a significant PPI effect (*t*-test, a cluster corrected FDR-threshold of $P < 0.05$) is overlaid on an MNI single-subject T1 scan.

vmPFC (ROI-1, ROI-2, ROI-3 and ROI-4) and other areas in the brain using a PPI analysis (O'Reilly *et al.*, 2012). The interaction factor was defined as the element-by-element product of the (mean-centered) time course for (self *vs* high reward) condition and the (demeaned) seed ROI time course. Eight separate PPIs were performed [four seed regions (ROI-1, ROI-2, ROI-3, ROI-4) and two psychological factors of interest (self > high reward, high reward > self)].

These analyses showed that, compared to high reward, self-associated stimuli increased functional coupling between ROI-3 and two clusters in the right and left frontal pole (peak at $x = 21$, $y = 56$, $z = -04$ and $x = -30$, $y = 53$, $z = -02$, respectively). There was also increased coupling between ROI-4 and the left frontal pole (peak at $x = -21$, $y = 53$, $z = 04$), extending to the left and right superior frontal gyri (peak at $x = -21$, $y = 29$, $z = 52$ and $x = 21$, $y = 20$, $z = 40$, respectively) (Figure 6), for self *vs* high reward stimuli. In contrast, high reward, compared to self-stimuli increased functional coupling between ROI-3 and the left inferior temporal gyrus (peak at $x = -45$, $y = -20$, $z = -23$) and the left temporal pole (peak at $x = -51$, $y = 11$, $z = -26$). A PPI effect for high reward relative to the self was found for connectivity between ROI-4 and the left middle temporal gyrus (MTG) (peak at $x = -51$, $y = -20$, $z = -10$; Figure 6). However, neither the self nor high reward was significantly associated with differential functional connectivity between either ROI-1 or ROI-2 and the rest of the brain.

The relationship between changes in functional connectivity and behavioral performance

Previous studies have demonstrated that the strength of functional coupling between brain areas involved in self-referential and reward tasks is linked to individual differences in behavioral performance (Smith *et al.*, 2010; Sui and Humphreys, 2013; Smith *et al.*, 2014). Here we asked whether differences in functional connectivity for self and high reward stimuli related to the two behavioral biases. To test this, we correlated the differences in connectivity strength with RT biases for self and high reward using non-parametric correlations with a bias-corrected and accelerated (BCa) bootstrapping procedure. Specifically, we

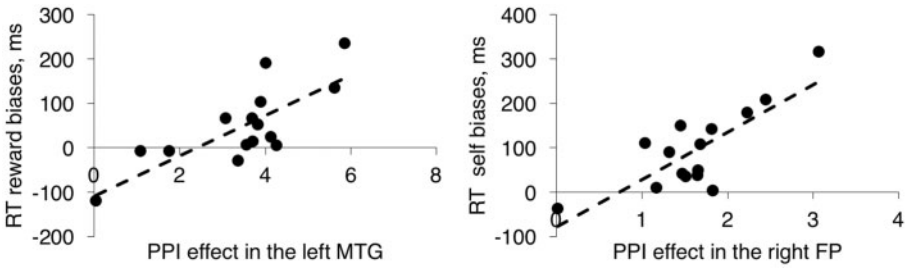

Fig. 8. Correlation graphs showing the relationship of behavioral biases in relation to the PPI effect for (A) self, in the right FP and (B) reward, in the left MTG. The individual PPI estimates across the voxel matrix were correlated with individual behavioral biases ($r_s = 0.61$, $P = 0.01$, BCa 95% CI [0.32, 0.9]; $r_s = 0.81$, $P < 0.001$, BCa 95% CI [0.52, 0.95]; for self and reward, respectively).

correlated RT biases with (i) the ROI-specific group maps showing increased connectivity for self (self > high reward) and (ii) the ROI-specific group maps of the increased connectivity for high reward (high reward > self). Clusters of voxels showing significant correlations were thresholded at $P = 0.005$ and corrected for multiple comparisons using FDR-correction ($P < 0.05$).

We observed two significant clusters of correlations: (i) between self RT biases and functional connectivity differences (self- > high reward) to the right frontal pole when ROI-3 was a seed and (ii) between RT reward-biases and connectivity differences (high reward > self) to the left MTG when ROI-4 was a seed (Figure 7). To demonstrate the direction of the correlation effects, individual PPI estimates in the right frontal polar (FP) and the left MTG were plotted against RT biases for self and high reward, respectively (Figure 8). Importantly, in our PPI models the psychological context of interest was simply presentation of self and high reward associations (i.e. unmodulated regressors) which were therefore independent of the behavioral responses. The finding that the PPI effects in the frontal pole for self and the MTG for high reward positively correlate with behavioral biases provides additional support for distinct neural mechanisms supporting self- and reward-biases in the posterior vmPFC.

Discussion

This study addressed the relationship between self and reward processing across the anterior–posterior axis in the vmPFC using an associative-matching procedure that allows direct comparison of performance biases for self and high reward (Sui *et al.*, 2012; Sui and Humphreys, 2013).

The results of our univariate ROI analyses showed a differential involvement of the vmPFC in self and reward tasks. Specifically, the anterior part of the vmPFC was more engaged in the reward task, while the posterior part responded more strongly in the self-task. Furthermore, the whole-brain univariate analysis provides further support for these results (see Supplementary Material). This finding is in line with meta-analyses on univariate data for reward (Liu *et al.*, 2011) and self-referential processing (Northoff *et al.*, 2006; Denny *et al.*, 2012) and indicates spatial segregation between self and reward within the vmPFC. However, the mean activation differences do not encode the relationship between voxels within an ROI that can be crucial for understanding brain mechanisms underlying biased performance for self and reward and their mutual relationship.

Our MVPA procedure went beyond these univariate analyses in demonstrating that all ROIs along the anterior–posterior axis of the vmPFC have strong representational content for self- and reward-biases. The patterns of activity associated with either self- or reward-biases yielded high classification accuracy indicating that each of our ROI's contains information about the biases. However, the similarity of activity patterns for self and reward linearly decreases from the anterior vmPFC toward the posterior part of the vmPFC, as indexed by a reduction in cross-task classification accuracy. Failing to generalize self to reward in the posterior vmPFC suggests functional dissimilarity in the processing of the biases. This finding can be interpreted as a gradual shift from monetary reward-biases to self-biases along the anterior–posterior axis of the vmPFC. Importantly, our analyses also demonstrated that this shift was specific to the vmPFC. When we assessed stimulus classification along an axis from the dmPFC to the avmPFC, we found no evidence for the successful classification of the stimuli.

Our PPI analysis provides further evidence for the functional dissociations between self- and reward-biases and helps to explain the difference in activation patterns between the anterior and posterior part of the vmPFC. Specifically, the pvmPFC regions, but not avmPFC regions, showed changes in functional connectivity to polar regions of the frontal lobe in relation to self-stimuli, and these correlated with self-biases in behavior. In contrast, stimuli related to high reward generate changes in functional connectivity from the pvmPFC to the middle temporal cortex and the strength of the connectivity linked to an individual's reward-bias. There is evidence that the lateral part of the FP cortex represents cognitive calculations of stimulus values (Bludau *et al.*, 2014) and the dorsal part is associated with action-related coding (e.g. stimulus–response mapping) (O'Reilly *et al.*, 2012; Orr *et al.*, 2015). Here the clusters of the significant PPI effect in the FP for self-associations showed peak activity at $x = 21$, $y = 56$, $z = 04$ (and $x = -30$, $y = 53$, $z = -02$), which corresponds to the dorsolateral part of the FP indicating that self-related stimuli may be associated with enhanced stimulus–response mapping. Furthermore, prior evidence (Gilbert *et al.*, 2010) indicates that the lateral part of the FP is coactivated with the anterior cingulate cortex and anterior insula—regions known to support self-referential and emotional processing (Northoff *et al.*, 2006; Denny *et al.*, 2012). Therefore, it is plausible to assume that the self may differentiate from reward by eliciting a greater emotional response (Ma and Han, 2010; Chavez *et al.*, 2016). For example, recent study using MVPA in a cross-domain neural population decoding paradigm directly tested the idea that self-referential thought elicits positive affect and suggested that this information can be decoded from activity in the posterior part of the vmPFC (Chavez *et al.*, 2016).

Prior studies showed that the pvmPFC responds differentially to social and monetary reward values (Smith *et al.*, 2014; Zaki *et al.*, 2014) and there is evidence that functional connectivity between the pvmPFC and the MTG increases with increasing social reward evaluation (Smith *et al.*, 2014). Our finding that

reward-biases modulate the connectivity between the pvmPFC and the MTG confirms the results of previous studies and provides further evidence of functional differences between self- and reward-biases in the pvmPFC.

Conclusion

Taken together our results make three important contributions to the debate about the relationship between self- and reward-biases in the vmPFC. First, the inferences that are made about this relationship depend on the level of information associated with neural responses for self- and reward-biases. In particular, overall activation levels (a mean activation difference) supports the spatial separation between processing of self- and reward-biases in the vmPFC and comply with the segregation model (Northoff and Hayes, 2011) where value assignment and self-referential assignment are regarded as different and spatially non-overlapping processes. However, analysis at the multi-voxel pattern level provides evidence for a complex relationship between self and reward along an anterior–posterior axis in the vmPFC, supporting the parallel processing model (Northoff and Hayes, 2011). Second, the finding that both self- and reward-biases have strong representations across the anterior–posterior axis indicates that this area is critical for processing self- and reward-related information. We speculate here that the representations may be tuned to a specific task which may explain consistency in reporting of activation in various tasks linked to personal and reward relevance along the anterior–posterior axis. Third, the shift from the specific classification of monetary reward to self is a characteristic of the vmPFC axis, and within the posterior sections of this axis, self-bias and monetary reward-bias are distinguished at the neural level.

Directions for further research

An important next step toward developing the models of the relationship between self and reward processing will be to examine whether and where other types of reward (e.g. social reward) (Wang *et al.*, 2016) relate to self-biases using both univariate and multivariate approaches. Exploring this direction will provide a better understanding of the interaction between self and reward and psychological functions associated with each concept and may have implications for clarifying a contributing cause of behavioral change in neurodegenerative diseases (e.g. Alzheimer's disease and frontotemporal dementia).

Beyond asking which brain regions showed selective responses for self and reward, the most important question to be answered is *how* class information (self and reward) is presented in the brain. Pattern characterization approach using linear and non-linear classifiers, perhaps, has the greatest potential to explore the complex combination of voxel activities for self and reward.

Supplementary data

Supplementary data are available at *SCAN* online.

Funding

The work was supported by grants from the European Research Council to G.H. (PePe 323883) and 'MRC Career Development Award' to M.S.

Conflict of interest. None declared.

References

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *Neuroimage*, **38**(1), 95–113.

Bäckström, M., Björklund, F. (2013). Social desirability in personality inventories: symptoms, diagnosis and prescribed cure. *Scandinavian Journal of Psychology*, **54**(2), 152–9.

Behrens, T. (2013) Neural mechanisms underlying human choice in the frontal cortex. In: *Neurosciences and the Human Person: New Perspectives on Human Activities Pontifical Academy of Sciences*, Scripta Varia 121, Vatican City, 1–13. Available: www.casinapioiv.va/content/dam/accademia/pdf/sv121/sv121-behrens.pdf

Bludau, S., Eickhoff, S.B., Mohlberg, H., *et al.* (2014). Cytoarchitecture, probability maps and functions of the human frontal pole. *Neuroimage*, **93(Pt 2)**, 260–75.

Burgess, P.W., Gilbert, S.J., Dumontheil, I. (2007). Function and localization within rostral prefrontal cortex (area 10). *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **362**(1481), 887–99.

Chavez, R.S., Heatherton, T.F., Wagner, D.D. (2016). Neural population decoding reveals the intrinsic positivity of the self. *Cerebral Cortex*, **27**(11), 5222–5229.

Christoff, K., Cosmelli, D., Legrand, D., Thompson, E. (2011). Specifying the self for cognitive neuroscience. *Trends in Cognitive Sciences*, **15**(3), 104–12.

Clithero, J.A., Rangel, A. (2014). Informatic parcellation of the network involved in the computation of subjective value. *Social Cognitive and Affective Neuroscience*, **9**(9), 1289–302.

Coutanche, M.N. (2013). Distinguishing multi-voxel patterns and mean activation: why, how, and what does it tell us? *Cognitive, Affective and Behavioral Neuroscience (CABN)*, **13**(3), 667–73.

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), 1742–52.

Fossati, P., Hevenor, S.J., Lepage, M., *et al.* (2004). Distributed self in episodic memory: neural correlates of successful retrieval of self-encoded positive and negative personality traits. *Neuroimage*, **22**(4), 1596–604.

Frings, C. and Wentura, D. (2014). Self-prioritization processes in action and perception. *J Exp Psychol Hum Percept Perform*, **40**, 1737–1740.

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

Gilbert, S.J., Gonen-Yaacovi, G., Benoit, R.G., Volle, E., Burgess, P.W. (2010). Distinct functional connectivity associated with lateral versus medial rostral prefrontal cortex: a meta-analysis. *Neuroimage*, **53**(4), 1359–67.

Han, S., Northoff, G. (2008). Culture-sensitive neural substrates of human cognition: a transcultural neuroimaging approach. *Nature Reviews. Neuroscience*, **9**(8), 646–54.

Haxby, J.V. (2012). Multivariate pattern analysis of fMRI: the early beginnings. *Neuroimage*, **62**(2), 852–5.

Haxby, J.V., Connolly, A.C., Guntupalli, J.S. (2014). Decoding neural representational spaces using multivariate pattern analysis. *Annual Review of Neuroscience*, **37**(1), 435–56.

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**, 2425–30.

Kable, J.W., Glimcher, P.W. (2007). The neural correlates of subjective value during intertemporal choice. *Nature Neuroscience*, **10**(12), 1625–33.

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), 785–94.

Knutson, B., Taylor, J., Kaufman, M., Peterson, R., Glover, G. (2005). Distributed neural representation of expected value. *Journal of Neuroscience*, **25**(19), 4806–12.

Konstabel, K., Aavik, T., Allik, J. (2006). Social desirability and consensual validity of personality traits. *European Journal of Personality*, **20**(7), 549–66.

Kriegeskorte, N., Mur, M., Bandettini, P. (2008) Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2, doi: 10.3389/neuro.06.004.2008

Liu, X., Hairston, J., Schrier, M., Fan, J. (2011). Common and distinct networks underlying valence and processing stages: a meta-analysis of functional neuroimaging studies. *Neuroscience and Biobehavioral Reviews*, **35**(5), 1219–36.

Ma, Y., Han, S. (2010). Why we respond faster to the self than to others? An implicit positive association theory of self-advantage during implicit face recognition. *Journal of Experimental Psychology: Human Perception and Performance*, **36**(3), 619–33.

Macrae, C.N., Visokomogilski, A., Golubickis, M., Cunningham, W.A., Sahraie, A. (2017). Self-relevance prioritizes access to visual awareness. *Journal of Experimental Psychology: Human Perception and Performance*, **43**(3), 438–43.

Mitchell, J.P., Macrae, C.N., Banaji, M.R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron*, **50**(4), 655–63.

Moran, J.M., Macrae, C.N., Heatherton, T.F., Wyland, C.L., Kelley, W.M. (2006). Neuroanatomical evidence for distinct cognitive and affective components of self. *Journal of Cognitive Neuroscience*, **18**(9), 1586–94.

Nelissen, N., Stokes, M., Nobre, A.C., Rushworth, M.F.S. (2013). Frontal and parietal cortical interactions with distributed visual representations during selective attention and action selection. *Journal of Neuroscience*, **33**(42), 16443–58.

Nicolle, A., Klein-Flügge, M.C., Hunt, L.T., Vlaev, I., Dolan, R.J., Behrens, T.E. (2012). An agent independent axis for executed and modeled choice in medial prefrontal *cortex*. *Neuron*, **75**(6), 1114–21.

Northoff, G. (2015). 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. *Cognitive Neuroscience*, **1**(4), 203–222.

Northoff, G., Hayes, D.J. (2011). Is our self nothing but reward? *Biological Psychiatry*, **69**(11), 1019–25.

Northoff, G., Heinzel, A., De Greck, M., Bermpohl, F., Dobrowolny, H., Panksepp, J. (2006). Self-referential processing in our brain – a meta-analysis of imaging studies on the self. *Neuroimage*, **31**(1), 440–57.

O'Reilly, J.X., Woolrich, M.W., Behrens, T.E., Smith, S.M., Johansen-Berg, H. (2012). Tools of the trade: psychophysiological interactions and functional connectivity. *Social Cognitive and Affective Neuroscience*, **7**(5), 604–9.

Ojala, M., Garriga, G. (2010). Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, **11**, 1833–63.

Orr, J.M., Smolker, H.R., Banich, M.T. (2015). Organization of the human frontal pole revealed by large-scale DTI-based connectivity: implications for control of behavior. *PLoS One*, **10**(5), e0124797.

Padoa-Schioppa, C., Assad, J.A. (2008). The representation of economic value in the orbitofrontal cortex is invariant for changes of menu. *Nature Neuroscience*, **11**(1), 95–102.

Roy, M., Shohamy, D., Wager, T.D. (2012). Ventromedial prefrontal-subcortical systems and the generation of affective meaning. *Trends in Cognitive Sciences*, **16**(3), 147–56.

Rushworth, M.F., Noonan, M.P., Boorman, E.D., Walton, M.E., Behrens, T.E. (2011). Frontal cortex and reward-guided learning and decision-making. *Neuron*, **70**(6), 1054–69.

Schultz, W., Tremblay, L., Hollerman, J.R. (2000). Reward processing in primate orbitofrontal cortex and basal ganglia. *Cerebral Cortex*, **10**(3), 272–84.

Smith, D.V., Clithero, J.A., Boltuck, S.E., Huettel, S.A. (2014). Functional connectivity with ventromedial prefrontal cortex reflects subjective value for social rewards. *Social Cognitive and Affective Neuroscience*, **9**(12), 2017–25.

Smith, D.V., Hayden, B.Y., Truong, T.K., Song, A.W., Platt, M.L., Huettel, S.A. (2010). Distinct value signals in anterior and posterior ventromedial prefrontal cortex. *Journal of Neuroscience*, **30**(7), 2490–5.

Spaak, E., Watanabe, K., Funahashi, S., Stokes, M.G. (2017). Stable and dynamic coding for working memory in primate prefrontal cortex. *Journal of Neuroscience*, doi: 10.1523/JNEUROSCI.3364-16.2017.

Stokes, M., Thompson, R., Cusack, R., Duncan, J. (2009). Top-down activation of shape-specific population codes in visual cortex during mental imagery. *Journal of Neuroscience*, **29**(5), 1565–72.

Stokes, M.G. (2015). "Activity-silent" working memory in prefrontal cortex: a dynamic coding framework. *Trends in Cognitive Sciences*, **19**(7), 394–405.

Sui, J., He, X., Humphreys, G.W. (2012). Perceptual effects of social salience: evidence from self-prioritization effects on perceptual matching. *Journal of Experimental Psychology: Human Perception and Performance*, **38**(5), 1105–17.

Sui, J., Humphreys, G.W. (2013). Self-referential processing is distinct from semantic elaboration: evidence from long-term memory effects in a patient with amnesia and semantic impairments. *Neuropsychologia*, **51**(13), 2663–73.

Sui, J., Rotshtein, P., Humphreys, G.W. (2013). Coupling social attention to the self forms a network for personal significance. *Proceedings of the National Academy of Sciences of the United States of America*, **110**(19), 7607–12.

Sui, J., Sun, Y., Peng, K., Humphreys, G.W. (2014). The automatic and the expected self: separating self- and familiarity biases effects by manipulating stimulus probability. *Attention, Perception and Psychophysics*, **76**(4), 1176–84.

Sui, J., Yankouskaya, A., Humphreys, G.W. (2015). Super-capacity me! Super-capacity and violations of race independence for self- but not for reward-associated stimuli. *Journal of Experimental Psychology: Human Perception and Performance*, **41**(2), 441–52.

Sui, J., Zhu, Y., Han, S. (2006). Self-face recognition in attended and unattended conditions: an event-related brain potential study. *Neuroreport*, **17**(4), 423–7.

Sun, Y., Fuentes, L.J., Humphreys, G.W., Sui, J. (2016). Try to see it my way: domain-specific embodiment enhances self and friend-biases in perceptual matching. *Cognition*, **153**, 108–17.

Tamir, D.I., Mitchell, J.P. (2012). Disclosing information about the self is intrinsically rewarding. *Proceedings of the National Academy of Sciences of the United States of America*, **109**(21), 8038–43.

van Buuren, M., Gladwin, T.E., Zandbelt, B.B., Kahn, R.S., Vink, M. (2010). Reduced functional coupling in the default-mode network during self-referential processing. *Human Brain Mapping*, **31**(8), 1117–27.

Wang, K.S., Smith, D.V., Delgado, M.R. (2016). Using fMRI to study reward processing in humans: past, present, and future. *Journal of Neurophysiology*, **115**, 1664–78.

Whitfield-Gabrieli, S., Nieto-Castanon, A. (2012). Conn: a functional connectivity toolbox for correlated and anticorrelated brain networks. *Brain Connectivity*, **2**(3), 125–41.

Williams, M.A., Dang, S., Kanwisher, N.G. (2007). Only some spatial patterns of fMRI response are read out in task performance. *Nature Neuroscience*, **10**(6), 685–6.

Zaki, J., Lopez, G., Mitchell, J.P. (2014). Activity in ventromedial prefrontal cortex co-varies with revealed social preferences: evidence for person-invariant value. *Social Cognitive and Affective Neuroscience*, **9**(4), 464–9.