Cerebral Cortex, November 2017;27: 5222–5229

doi: 10.1093/cercor/bhw302
Advance Access Publication Date: 24 September 2016
Original Article

ORIGINAL ARTICLE

Neural Population Decoding Reveals the Intrinsic Positivity of the Self

Robert S. Chavez[1], Todd F. Heatherton[2], and Dylan D. Wagner[1]

[1]The Ohio State University, Department of Psychology, Columbus, OH 43210, USA and [2]Dartmouth College, Department of Psychological and Brain Sciences, Hanover, NH 03755, USA

Address correspondence to Robert S. Chavez, The Ohio State University, Department of Psychology, Lazenby Hall, 1827 Neil Avenue, Columbus, OH 43210, USA. Email: chavez.95@osu.edu

Abstract

People are motivated to hold favorable views of themselves, which manifests as a positivity bias when evaluating their own performance and abilities. However, it remains an open question whether positive affect is an essential component of people's self-concept. Prior functional neuroimaging research demonstrated that similar regions of the brain support positive affect and self-referential processing, although a direct test of their shared representation has yet to be examined. Here we use functional magnetic resonance imaging in conjunction with multivariate pattern analysis in a cross-domain neural population decoding paradigm. We found that a multivariate pattern classifier model trained to dissociate neural responses to viewing positively and negatively valenced images can dissociate thinking about oneself from a close friend during a lexical trait-judgment task commonly used in the study of self-referential processing. Cross-domain classification accuracy was found to be highest in the ventral medial prefrontal cortex (vMPFC), a region previously implicated in both self-referential processing and positive affect. These results show that brain responses during self-referential processing can be decoded from multi-voxel activation patterns in the vMPFC when viewing positively valenced material, thereby providing evidence that positive affect may be a central component of the mental representation of the self.

Key words: functional magnetic resonance imaging, medial prefrontal cortex, multivariate pattern analysis, positive affect, self-representation

Introduction

The ability to generate a mental representation of oneself is a defining characteristic of human cognition. Unlike other species, humans have a seemingly unique ability to introspect and evaluate their traits and behaviors and form conclusions about themselves (Bem 1967). However, evaluative attitudes of the self are not always objective. In general, people are motivated to emphasize their positive attributes (Taylor and Brown 1988) and deemphasize their less favorable ones (Dunning and Cohen 1992). Indeed, researchers have demonstrated a broadly held positivity bias in evaluations of the self (Baumeister et al. 1989), so much so that its absence is often associated with psychological disorders, such as depression and anxiety (Lewinsohn et al. 1980; Moore and Fresco 2012). This long tradition of research showing a positivity bias for self raises the possibility that the self-concept is fundamentally linked to positive affect.

Functional neuroimaging studies offer some initial evidence that the self and positive affect share a common neural representation, in that studies of self-referential processing typically recruit brain regions that overlap with those examining the affective processing of positive information. Indeed, meta-analyses have shown that self-referential processing is most consistently associated with activation of the ventral medial prefrontal cortex (vMPFC) (Denny et al. 2012; Wagner et al. 2012), an area also involved in positive evaluation and reward value (Liu et al. 2011). Similarly, subcortical structures typically associated with positive affect and reward-related

processing—such as the ventral striatum—have also been shown to preferentially activate to self-relevant information (Phan et al. 2004; Denny et al. 2012) and respond when disclosing information about the self to others (Tamir and Mitchell 2012). This shared neural circuitry has led some researchers to propose that the representation of the self may simply be a special case of reward processing (Northoff and Hayes 2011). Studies testing this hypothesis have found that both neural responses to thinking about the self and winning gambles share common areas in reward system, including the MPFC, ventral striatum, and ventral tegmental area (de Greck et al. 2008; Enzi et al. 2009).

However, simply because two processes activate overlapping regions of the brain does not necessarily imply that they share a common underlying cognitive mechanism. For example, researchers have found that physical pain and social rejection both show activation of the dorsal anterior cingulate cortex (dACC) using standard univariate functional magnetic resonance imaging (fMRI) analyses (Eisenberger et al. 2003; Kross et al. 2011). However, multivariate fMRI decoding methods have shown that the pattern of neural responses within the dACC can nevertheless distinguish between the two processes (Woo et al. 2014). Thus, information about recruitment of a particular cognitive process was better reflected in distributed multi-voxel activation patterns than in the aggregated univariate response of those voxels.

Indeed, using multivariate pattern analysis (MVPA), researchers have demonstrated that information about particular cognitive processes can be decoded across different domains and stimulus modalities. For example, one study (Parkinson et al. 2014) found that activation in the right inferior parietal lobule reflected a common computation of distance across the spatial, temporal, and social domains. In another study (Chikazoe et al. 2014), researchers demonstrated that the vMPFC and medial orbitofrontal cortex use a common neural representation of valence when processing both visual and gustatory sensory information. Given the results of these studies, cross-domain multivariate decoding methods may offer a more direct approach for investigating the shared representation of self-representation and positive affect than traditional univariate approaches.

The aim of the current study is to test the hypothesis that self-referential thought and positive affect share an underlying neural representation. Specifically, we hypothesized that positive affect is intrinsically elicited by self-referential thought and this response can be decoded from brain responses patterns to affective processing in an independently measured domain. Using a cross-domain MVPA decoding approach, a classifier was trained to dissociate neural responses viewing positively and negatively valenced images in the affective domain. This was then used to predict brain responses when participants were thinking about the self versus thinking about a close friend in a lexical trait-judgment task commonly employed in studies of self-referential processing (Kelley et al. 2002). To the extent that the classifier model built on affective information can accurately classify neural responses during self-referential processing, this would provide evidence that the self and positive affect share a common neural representation.

Materials and Methods

Participants

Fourteen right-handed subjects (7 female) between the ages of 20 and 33 participated in this study. All subjects were screened for MRI contraindications, had normal or corrected-to-normal vision, were native English speakers, and reported no history of psychiatric or neurological conditions. Subjects underwent an MRI protocol which included 10 functional runs and a high-resolution anatomical scan. Subjects gave informed consent in accordance with the guidelines set by the Committee for the Protection of Human Subjects at Dartmouth College and were paid for their participation.

Image Acquisition

Magnetic resonance imaging was conducted with a Philips Achieva 3.0 Tesla scanner using a 32-channel phased array coil. Structural images were acquired using a T1-weighted MP-RAGE protocol (220 sagittal slices; TR: 8.176 ms; TE: 3.72 ms; flip angle: 8°; 1 mm isotropic voxels). Functional images were acquired using a T2*-weighted echo planar sequence (TR: 2000 ms; TE: 35 ms; flip angle: 90°). For each participant, 10 runs of 194 whole-brain volumes (35 axial slices per whole-brain volume, 3 mm × 3 mm × 3.486 mm voxel size) were collected. Acquisition parameters for both functional tasks were identical. The total length of time of the entire scanning session was approximately 75 min.

Procedure

While being scanned, subjects completed tasks in two separate domains: a visual valence task and a lexical self/other trait-judgment task. Five runs of each task (10 runs total) were collected in an interleaved order and the stimuli within each domain were ordered randomly within each run. No two subjects in the analysis were presented with the same order of stimuli across the whole experiment. Both domains were implemented within a rapid event-related design.

For the visual domain task, subjects were shown images against a black background followed by a fixation cross. A total of 400 images (80 per run) were presented for 2000 ms each along with intermittent passive fixation trials of variable durations (2000–6000 ms) to introduce jitter into the fMRI time series. The images presented depicted a variety of positively and negatively valenced scenes that were matched for social content (i.e., the presence of people or animals in the images) and randomly ordered within each run. During the presentation of the images, subjects were asked to rate pictures via an fMRI compatible button box as being either positive or negative on a 1–4 scale (i.e., 1 = "Very negative"; 4 = "Very positive"). These ratings were then used to subsequently assign images to either a positive or negative affect category; in this way all participants received a subject-specific model based of their own subjective evaluations of valence for the images. To estimate neural responses to each valence category, pictures rated 1 or 2 were collapsed to reflect negatively valenced trials, whereas pictures rated 3 or 4 were collapsed to reflect positively valenced trials.

For the lexical domain task, we employed a version of the task by Heatherton and colleagues (Heatherton et al. 2006) used to distinguish thinking about oneself from a friend. Subjects were present with a screen containing two words stacked vertically and printed in white font on black background. During each trial, the top word displayed either "SELF" or "FRIEND" and the bottom word displayed 1 of 400 possible trait adjectives (80 per run) for 2000 ms along with intermittent passive fixation trials of variable durations (2000–6000 ms) for jitter. These words were selected from a list normed for valence (Anderson 1968)

such that half of the presented words were positively valenced (e.g., "nice", "competent") and the other half were negatively valenced (e.g., "boring", "lazy"). For trials in which the word presented on the top said "SELF", subjects were instructed to rate via button box how much the bottom word described them using the scale 1 ("not at all") through 4 ("very much"). For trials in which the word presented on the top said "FRIEND", subjects were asked to rate the bottom word for how much it described a close friend using the same scale. Before scanning, subjects were asked to think of a best friend or close friend and use that same person when making friend judgments throughout the whole experiment. Neural responses were then estimated separately for each trial type.

Image Analysis

Preprocessing

The fMRI data were preprocessed and voxel responses were estimated using FSL (http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/) (Smith et al. 2004). First, all slices were interpolated to a common time point (i.e., slice-time correction) to correct for differences in slice acquisition. We used mean-based intensity normalization of all volumes by the same factor and high-pass temporal filtering (Gaussian-weighted least-squares straight line fitting, with sigma = 90.0 s). Time-series statistical analyses were carried out using local autocorrelation correction. Consistent with previous MVPA studies, no spatial smoothing was applied in order to capitalize on the fine-grained information within the multivariate voxel patterns (Haxby et al. 2001; Norman et al. 2006). All first level analyses were performed in native fMRI space for each run before being aligned to the middle functional run of the experiment for each participant. Normalized (i.e., z-score) voxel responses for each stimulus category within each run were submitted to the MVPA classification analysis, as these measurements have been shown to outperform raw beta-weight values when using general linear model estimates from rapid event-related designs (Misaki et al. 2010).

Classification Analyses

Cross-domain classification analyses were performed using PyMVPA (Hanke et al. 2009). Using a searchlight procedure to test MVPA classification performance across the whole brain (Kriegeskorte et al. 2006), a 9 mm radius spherical region of interest (ROI) was moved throughout each subject's brain to perform cross-domain decoding for each subject. Within each searchlight ROI, a linear support vector machine (SVM) with fixed regularization parameter of $C = -1$ was trained to discriminate local patterns of fMRI responses to positive and negative images from the five affective domain runs. The SVM decision boundary trained on valence was then tested on response patterns from the five lexical task runs to discriminate between self and friend trials. This process was then reversed by training on the self/friend domain then testing in the affective domain. The result of these two cross-domain classification analyses was averaged to produce the classification accuracy. Classification accuracy was determined by the percentage of runs in which neural responses to self trials were classified as positive valence trials and responses to friend trials were classified as negative valence trials (and vice versa for the reverse training procedure). Given that people tend to view close friends positively, this approach is a relatively conservative test of the classification accuracy of positive valence and self-referential processing. Classification accuracy was computed across data folds within each searchlight ROI, resulting in a whole brain map of percent accuracy scores at each voxel for each participant.

Group-level Analysis

In order to test the correspondence across individuals, classification accuracy maps in each subject's native space were then registered to the Montreal Neurological Institute (MNI) standard stereotaxic space. A two-step normalization process was performed by aligning functional data to each subject's anatomical scan using linear registrations with FSL's FLIRT before registering it to the MNI template using nonlinear registration with FSL's FNIRT. Because classification accuracies are not normally distributed, non-parametric permutation tests in FSL's Randomize were used for voxel-wise statistical inferences (Winkler et al. 2014). Classification accuracies in each voxel were tested against chance performance (50%) using an exhaustive one-sample permutation t-test. Threshold-free cluster enhancement (TFCE; Smith and Nichols 2009) was used to identify significant clusters from the permutation tests. The final results were thresholded at a TFCE cluster-corrected P-value < 0.05 after controlling for family-wise error rate.

Representational Similarity Analysis

To more completely characterize the relationship among the conditions within each domain, we also implemented a representational similarity analysis (RSA; Kriegeskorte et al. 2008). RSA provides a method by which to measure the similarity between multi-voxel activity patterns from separate conditions by expressing the similarity between categories as a measure of distance. In order to identify regions in which to perform follow-up RSA, we first created regions of interest from clusters demonstrating significant classification accuracy at the group-level. These group-derived ROIs were then projected back into each subject's native fMRI space by inverting the transformations used in the standard space registration procedure. Voxel activation patterns were extracted within these masks for each condition across each run and correlated with one another in order to create a dissimilarity matrix of correlation distances between each pair of conditions. Dissimilarity matrices were then averaged across subjects to estimate and visualize the group-wise similarity structure of neural responses to each condition.

Results

Behavioral Results

Subjects completed five runs each of two different tasks (Fig. 1). The first was a visual affect task in which participants rated the valence of a series of images. The other task a lexical self/other trait-judgment task in which participants rated the degree to which a presented adjective described either themselves or their close friend. In order to successfully implement cross-domain decoding procedure using information from the affective domain, it is imperative that there is concurrence among the subjects about the valence of the visual images. Behavioral responses from the affective domain showed that valence ratings of the images had high inter-subject reliability [interclass correlation coefficient = 0.71 with a 95% confidence interval from 0.68 to 0.75], indicating that there was strong agreement among subjects about the valence of the images used in this task. For the trait-judgment task, there was a significant main effect of valence on the endorsement ratings of traits [$F(1,13) = 60.99$, $P < 0.0001$], indicating that subjects were

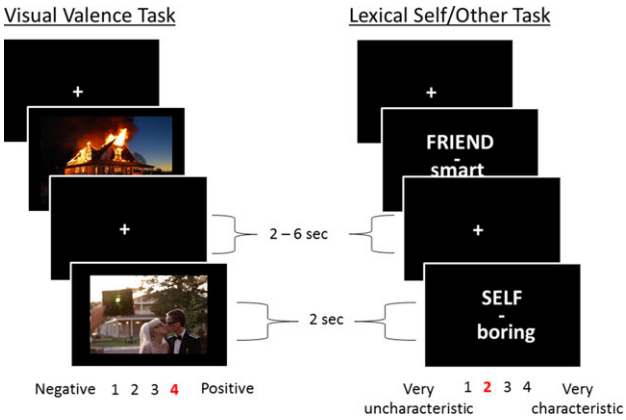

Figure 1. A schematic of the two tasks used in the current study. On the left is the affective task in which subjects rated the valence of a variety of images. On the right is the lexical task in which subjects made trait judgments of how much the presented word described either themselves or a close friend. Five runs of each task were collected for use in the classification analysis.

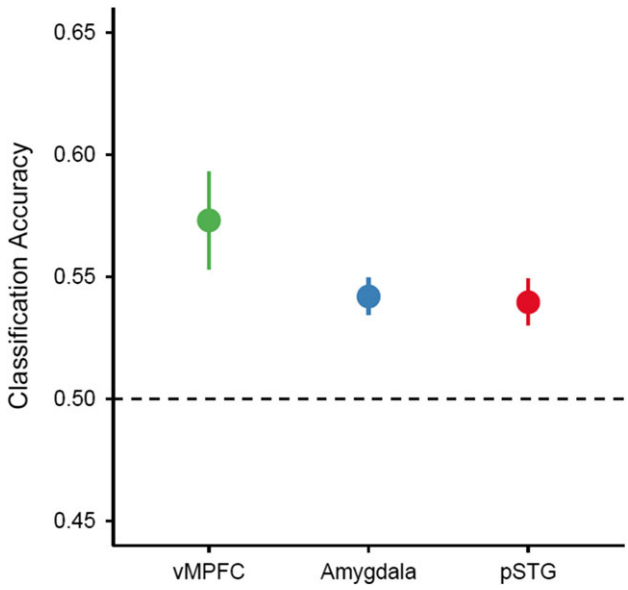

Figure 2. Classification accuracies within each significant ROI for the cross-domain decoding analysis across subjects. The vMPFC showed the highest classification accuracy among the three significant regions. The dashed line represents chance-level classification performance. Error bars represent the standard error of the mean.

more likely to endorse positive traits for both themselves and their friend. There was also a significant main effect of self/friend [$F(1, 13) = 6.11$, $P = 0.028$], such that subjects were more likely to endorse traits for themselves irrespective of valence. This effect appears to be driven by a greater endorsement for negative trait words for the self, although this interaction was not significant [$F(1,13) = 3.61$, $P = 0.08$]. Importantly, these results show that subjects were more willing to endorse negative traits for themselves than they were for their friend (see Supplementary Fig. 1). Thus, the accuracy classifier to decode self-relevant activity from positive affect in the fMRI data must be robust to this observed behavioral inclination to endorse more negatively valenced trait words for the self.

MVPA Results

Cross-domain decoding analyses were performed using a linear SVM classifier across the whole-brain with an iterative searchlight procedure (Kriegeskorte et al. 2006). The SVM decision boundary trained to dissociate viewing positively and negatively valenced images in the affective domain was then tested on each run of data from the self/other domain. This procedure was then reversed (i.e. trained to dissociate self and friend and tested on the affective domain) to produce the total accuracy scores from the cross-domain decoding procedure. The results of this analysis were consistent with the hypothesized role of the vMPFC in reflecting both positive affect and self-representation. The vMPFC (MNI: 8, 60, −16) showed the highest classification accuracy [$M = 57.31\%$, $SEM = 2.00\%$] when using positive and negative affect to dissociate thinking about the self versus a friend (Fig. 2). However, because classification accuracies may be driven either by high similarity between self and positive valence or high similarity between friend and negative valence, an RSA (Kriegeskorte et al. 2008) was used to better characterize the underlying relationship between these categories. Within this region of the vMPFC, the RSA results show the highest degree of similarity between each of the conditions was for positive affect and self trials (Fig. 3). This indicates that the classification accuracy of this region was largely driven by the similarity of multi-voxel neural responses between self and positive affect. There were no regions of the brain that showed the opposite effect (i.e., where self-referential brain activity was decoded from negative affect above chance).

In addition to the vMPFC, there were two other regions showing above chance decoding accuracy: the left amygdala (MNI: −12, −6, −18) [$M = 54.19\%$, $SEM = 0.75\%$], and left posterior superior temporal gyrus (pSTG; MNI: −56, −42, 22) [$M = 53.95\%$, $SEM = 0.95\%$]. RSA results indicate that classification accuracy in the amygdala was driven by the similarity between responses to a close friend and negative stimuli (Supplementary Fig. 2), whereas in the pSTG classification reflected the similarity between responses to self and positive images (Supplementary Fig. 3).

Discussion

Decades of research in psychology have shown that individuals are highly motivated to hold positive views about themselves. The current study provides evidence that self-referential thought elicits a similar pattern of neural responses to positive affect. Using cross-domain MVPA, we demonstrated that, in the vMPFC, neural responses to thinking about oneself versus a close friend in a trait-judgment task could be decoded using a model built from activation patterns elicited by positively and negatively valenced images in a separate domain. Subsequent analyses using RSA showed that classification performance within the vMPFC cluster was primarily driven by the similarity in response patterns between thinking of the self and viewing positive images, consistent with the hypothesis that this region supports a representation of the self that is positively valenced.

The vMPFC is implicated in a variety of reward-related tasks and shows preferential responding to rewarding outcomes (Knutson et al. 2003). In animal models using rats and primates, neural responses in the vMPFC have been shown to contribute to the assessment of value of both offered and chosen goods (Padoa-Schioppa and Assad 2006), the pursuit of outcome-specific rewards (Burke et al. 2008), and focal damage to this area has been found to impact valuation updating particularly

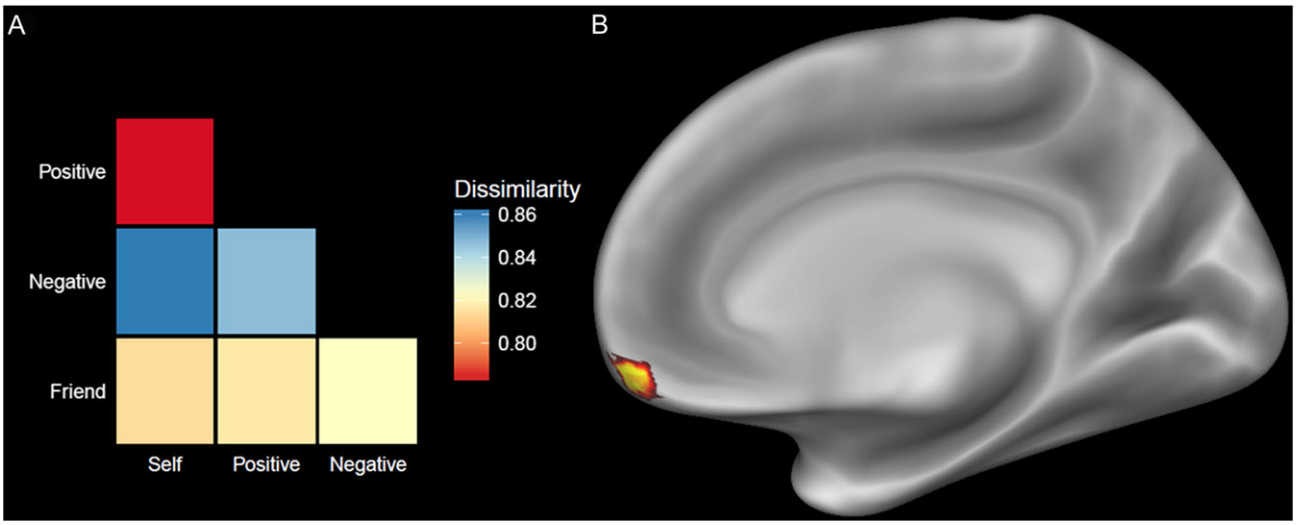

Figure 3. (A) Dissimilarity matrix from the results of the RSA, illustrating the relationship among patterns of neural responses to each condition within the vMPFC. Dissimilarity was measured by the correlation distance of voxel response patterns to each condition, which revealed that the classification accuracy within this region was driven by the similarity of neural responses to self and positive affect conditions. (B) Results from the whole-brain searchlight analysis showing the region of the vMPFC which above chance classification accuracy across subjects using non-parametric permutation tests.

(Rudebeck et al. 2013). In humans, fMRI activation patterns within this area has also been shown to represent both the anticipation and receipt of monetary rewards using MVPA methods similar to the current study (Kahnt et al. 2010). The current study aimed to link these findings with the body of work showing that the vMPFC is a critical region for supporting self-referential thought (Denny et al. 2012; Wagner et al. 2012) by showing that patterns of neural activity associated with self-referential thought can be decoded from activity patterns elicited by positively valenced stimuli in the affective domain. This provides support for the theoretical views that self-representation may reflect a particular case of reward-related cognition (Northoff and Hayes 2011) and that the vMPFC supports the integration of both valuation processes and self-referential thought (D'Argembeau 2013). More specifically, this finding extends previous work implicating the role of this region in processing both self-representation (Denny et al. 2012; Wagner et al. 2012) and positive affect (Knutson et al. 2001; Lindquist et al. 2016) by suggesting that these processes share a common cognitive representation beyond a mere anatomical overlap in the brain. More broadly, the results of the current study are also consistent with views of the functional role of the vMPFC as a hub supporting the integration of emotional and higher-level cognition (Chavez and Heatherton 2015a) which may be essential for dissociating conceptually meaningful affective representations from more basic ones (Roy et al. 2012).

The current findings also add to a growing body of work suggesting that evaluative self-referential processing is supported by the function and structure of frontostriatal reward networks. Previous work has shown that the vMPFC is involved in processing the self-relevance of emotional salience (Phan et al. 2004), the subjective importance of self-relevant information (D'Argembeau et al. 2012), and correlated with individual differences in the propensity for self-enhancement following negative social evaluation (Hughes and Beer 2013). Other work has shown that increases in self-relevance (Phan et al. 2004) and sharing information about the self (Tamir and Mitchell 2012) recruit the ventral striatum. Moreover, greater connectivity of vMPFC and ventral striatum has been shown when individuals make flattering evaluations of themselves (Flagan and Beer 2013), when individuals think they are being evaluated by their peers (Somerville et al. 2013), and individual differences in self-esteem are reflected in both the functional and structural connectivity of these regions (Chavez and Heatherton 2015b). Whereas these previous studies inferred a link between positive affect and self-relevant processing, the current study offers a more direct test of the idea that self-referential thought elicits positive affect and suggests that this information can be decoded from activity in the vMPFC specifically.

We interpret the results of the current study as providing evidence that positive affect is a key component to self-referential thought. However, an alternative interpretation would suggest the opposite direction: self-referential processing may be a central component of positive affect. Although this is a possibility, the design of the current study does not allow us to directly tease these two interpretations apart. However, our original interpretation is based on the notion that evolution shaped a complex mental representation of the self that emerges from the interactions among more basic mental processes (Barrett 2013) in which positive affect plays a role. In other words, self-representation may require affective processing, but not the other way around. Evidence for this directionality comes from studies showing that both the phylogenetic and ontogenetic development of affective responses precede the development of self-awareness. For example, immediately after birth, newborn infants show canonical responses to pleasant and unpleasant tastes that are consistent to those of closely related primates (Steiner et al. 2001). However, the development of self-awareness in infants does not to take place until between 15 and 22 months on average (Amsterdam 1972). Similarly, although there is evidence that many animals experience hedonic response to a variety of pleasant and unpleasant stimuli, only a handful of species have shown to reliably pass classic tests of self-awareness (e.g., the "mirror test"; Gallup 1970; Anderson and Gallup 2015). From these evolutionary and developmental perspectives, self-awareness is not necessary for the experience of positive affect, but the ability to

experience positive affect may still be necessary in order to maintain a normal representation of the self. We believe this speaks to the interpretation of the current results.

Although the self-positivity bias is well established, there are both adaptive and maladaptive consequences for holding such views. Despite their inherent bias, positive illusions of the self have been shown to facilitate adaptive responses to stressful situations and promote positive mental health outcomes (Taylor and Brown 1988; Taylor et al. 2008). Moreover, individuals with high self-esteem show increased resistance to affective disorders as well as greater initiative in the face of challenges (Baumeister et al. 2003). On the other hand, it has also been shown that students who engage in disingenuous self-enhancement score higher on measures of narcissism and show decreasing levels of self-esteem and less academic engagement over time (Robins and Beer 2001). Recent neuroimaging studies have begun to shed light on these issues by showing differential relationships self-esteem development and narcissism in their relationship to these neural systems. Whereas narcissism has shown an inverse relationship of structural connectivity of the vMPFC and ventral striatum (Chester et al. forthcoming), self-esteem shows a positive relationship to both the functional and structural connectivity of these systems (Chavez and Heatherton 2015b) which predict changes in self-esteem longitudinally (Chavez and Heatherton forthcoming). The current results can be seen as adaptive insofar as they reflect the processes involved in maintaining typical positive self-esteem rather than the hyper-defensive or self-aggrandizing processes associated with narcissism. Additional studies, however, will be needed to directly tease these possibilities apart.

Though the vMPFC had the highest classification accuracy in our analysis, the dorsomedial amygdala and left pSTG also showed above chance classification accuracy. This portion of the amygdala is responsible for output signals projecting to the brainstem and hypothalamus for triggering autonomic responses to such phenomena as threats, anxiety, or ambiguity (Davis and Whalen 2001), and our results may reflect responses to the uncertainty (Whalen 2007) of making these judgments about another person relative to better-known trait judgments of the self. The pSTG, on the other hand, is associated with processing the conjunction of visual and verbal information (Robins et al. 2009) and has been shown to reflect a supramodal representation of emotion across sensory modalities (Klasen et al. 2011), which is similar to the cross-domain approach employed in the present study. However, these findings were unexpected and the interpretation of our results for both of these regions remains speculative.

There are some potential caveats of current approach that warrant discussion. First, it should be noted that each of the trait adjectives used in the lexical self/friend task have an emotional valence which may raise concern that this would influence the decoding procedure. However, the valence categories of the traits during this task were matched for valence between self and friend trials. Moreover, the results from subjects' behavioral responses showed no significant difference in the number of positive traits endorsed for self trials than those for friend trials. As such, we believe that our analyses account for these issues and are consistent with dozens of studies that have previously used this task to probe self-referential processing in the brain (for examples see: Kelley et al. 2002; Heatherton et al. 2006; Krienen et al. 2010). Finally, although this procedure produced significant above-chance classification accuracy, the strength of this effect was modest and does not warrant an interpretation of one-to-one relationship of positive affect and self-referential processing. Indeed, there are several cognitive processes and phenomena that contribute to the mental representation of the self, including autobiographical memory, agency, social identity, and others. Relatedly, our results may only speak to these processes within healthy individuals and may not show similar responses within individuals diagnosed with affective disorders such as depression or anxiety. We believe that combining paradigms aimed at examining these related processes and populations together with information gleaned from this study will be an important avenue for future research.

In summary, the current study provides evidence that elements of self-referential processing can be decoded based on multivariate brain activity patterns to positive affect. This study also adds to a growing body of literature linking self-representation to the valuation and reward systems and suggests that positive affect may be a central psychological component in the human sense of self.

Supplementary Material

Supplementary material can be found at: http://www.cercor.oxfordjournals.org/.

Funding

National Institutes of Health (MH059282 to T.F.H.) and a National Science Foundation Graduate Research Fellowship (to R.S.C).

Notes

The authors thank Courtney Rogers for help with data collection and Bill Kelley and Thalia Wheatley for advice on the design of the study. *Conflict of Interest*: None declared.

References

Amsterdam B. 1972. Mirror self-image reactions before age two. Dev Psychobiol. 5:297–305.

Anderson JR, Gallup GG. 2015. Mirror self-recognition: a review and critique of attempts to promote and engineer self-recognition in primates. Primates. 56:317–326.

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

Barrett LF. 2013. Psychological construction: the Darwinian approach to the science of emotion. Emot Rev. 5:379–389.

Baumeister RF, Campbell JD, Krueger JI, Vohs KD. 2003. Does high self-esteem cause better performance, interpersonal success, happiness, or healthier lifestyles? Psychol Sci Public Interes. 4:1–44.

Baumeister RF, Tice DM, Hutton DG. 1989. Self-presentational motivations and personality differences in self-esteem. J Pers. 57:547–579.

Bem DJ. 1967. Self-perception: an alternative interpretation of cognitive dissonance phenomena. Psychol Rev. 74:183–200.

Burke KA, Franz TM, Miller DN, Schoenbaum G. 2008. The role of the orbitofrontal cortex in the pursuit of happiness and more specific rewards. Nature. 454:340–344.

Chavez RS, Heatherton TF. 2015a. Representational similarity of social and valence information in the medial pFC. J Cogn Neurosci. 27:73–82.

Chavez RS, Heatherton TF. 2015b. Multimodal frontostriatal connectivity underlies individual differences in self-esteem. Soc Cogn Affect Neurosci. 10:364–370.

Chavez RS, Heatherton TF. (in press). Structural integrity of frontostriatal connections predicts longitudinal changes in self-esteem. Soc Neurosci. doi:10.1080/17470919.2016.1164753.

Chester DS, Lynam DR, Powell DK, DeWall CN. Forthcoming. Narcissism is associated with weakened frontostriatal connectivity: a DTI study. Soc Cogn Affect Neurosci. 11:1036–1040.

Chikazoe J, Lee DH, Kriegeskorte N, Anderson AK. 2014. Population coding of affect across stimuli, modalities and individuals. Nat Neurosci. 17:1114–1122.

D'Argembeau A. 2013. On the role of the ventromedial prefrontal cortex in self-processing: the valuation hypothesis. Front Hum Neurosci. 7:372.

D'Argembeau A, Jedidi H, Balteau E, Bahri M, Phillips C, Salmon E. 2012. Valuing one's self: medial prefrontal involvement in epistemic and emotive investments in self-views. Cereb Cortex. 22:659–667.

Davis M, Whalen PJ. 2001. The amygdala: vigilance and emotion. Mol Psychiatry. 6:13–34.

de Greck M, Rotte M, Paus R, Moritz D, Thiemann R, Proesch U, Bruer U, Moerth S, Tempelmann C, Bogerts B, et al. 2008. Is our self based on reward? Self-relatedness recruits neural activity in the reward system. Neuroimage. 39:2066–2075.

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

Dunning D, Cohen GL. 1992. Egocentric definitions of traits and abilities in social judgment. J Pers Soc Psychol. 63:341–355.

Eisenberger NI, Lieberman MD, Williams KD. 2003. Does rejection hurt? An FMRI study of social exclusion. Science. 302: 290–292.

Enzi B, de Greck M, Prösch U, Tempelmann C, Northoff G. 2009. Is our self nothing but reward? Neuronal overlap and distinction between reward and personal relevance and its relation to human personality. PLoS One. 4:e8429.

Flagan T, Beer JS. 2013. Three ways in which midline regions contribute to self-evaluation. Front Hum Neurosci. 7:450.

Gallup GG. 1970. Chimpanzees: self-recognition. Science. 167: 86–87.

Hanke M, Halchenko YO, Sederberg PB, Hanson SJ, Haxby J V, Pollmann S. 2009. PyMVPA: a python toolbox for multivariate pattern analysis of fMRI data. Neuroinformatics. 7:37–53.

Haxby J V, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P. 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science. 293: 2425–2430.

Heatherton TF, Wyland CL, Macrae CN, Demos KE, Denny BT, Kelley WM. 2006. Medial prefrontal activity differentiates self from close others. Soc Cogn Affect Neurosci. 1:18–25.

Hughes BL, Beer JS. 2013. Protecting the self: the effect of social-evaluative threat on neural representations of self. J Cogn Neurosci. 25:613–622.

Kahnt T, Heinzle J, Park SQ, Haynes J-D. 2010. The neural code of reward anticipation in human orbitofrontal cortex. Proc Natl Acad Sci USA. 107:6010–6015.

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

Klasen M, Kenworthy CA, Mathiak KA, Kircher TTJ, Mathiak K. 2011. Supramodal representation of emotions. J Neurosci. 31:13635–13643.

Knutson B, Fong GW, Adams CM, Varner JL, Hommer D. 2001. Dissociation of reward anticipation and outcome with event-related fMRI. Neuroreport. 12:3683–3687.

Knutson B, Fong GW, Bennett SM, Adams CM, Hommer D. 2003. A region of mesial prefrontal cortex tracks monetarily rewarding outcomes: characterization with rapid event-related fMRI. Neuroimage. 18:263–272.

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

Kriegeskorte N, Mur M, Bandettini P. 2008. Representational similarity analysis – connecting the branches of systems neuroscience. Front Syst Neurosci. 2:4.

Krienen FM, Tu P-C, Buckner RL. 2010. Clan mentality: evidence that the medial prefrontal cortex responds to close others. J Neurosci. 30:13906–13915.

Kross E, Berman MG, Mischel W, Smith EE, Wager TD. 2011. Social rejection shares somatosensory representations with physical pain. Proc Natl Acad Sci USA. 108:6270–6275.

Lewinsohn PM, Mischel W, Chaplin W, Barton R. 1980. Social competence and depression: the role of illusory self-perceptions. J Abnorm Psychol. 89:203–212.

Lindquist KA, Satpute AB, Wager TD, Weber J, Barrett LF. 2016. The brain basis of positive and negative affect: evidence from a meta-analysis of the human neuroimaging literature. Cereb Cortex. 26:1910–1922.

Liu X, Hairston J, Schrier M, Fan J. 2011. Common and distinct networks underlying reward valence and processing stages: a meta-analysis of functional neuroimaging studies. Neurosci Biobehav Rev. 35:1219–1236.

Misaki M, Kim Y, Bandettini PA, Kriegeskorte N. 2010. Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. Neuroimage. 53: 103–118.

Moore MT, Fresco DM. 2012. Depressive realism: a meta-analytic review. Clin Psychol Rev. 32:496–509.

Norman K a, Polyn SM, Detre GJ, Haxby J V. 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn Sci. 10:424–430.

Northoff G, Hayes DJ. 2011. Is our self nothing but reward? Biol Psychiatry. 69:1019–1025.

Padoa-Schioppa C, Assad JA. 2006. Neurons in the orbitofrontal cortex encode economic value. Nature. 441:223–226.

Parkinson C, Liu S, Wheatley T. 2014. A common cortical metric for spatial, temporal, and social distance. J Neurosci. 34: 1979–1987.

Phan KL, Taylor SF, Welsh RC, Ho S-H, Britton JC, Liberzon I. 2004. Neural correlates of individual ratings of emotional salience: a trial-related fMRI study. Neuroimage. 21:768–780.

Robins DL, Hunyadi E, Schultz RT. 2009. Superior temporal activation in response to dynamic audio-visual emotional cues. Brain Cogn. 69:269–278.

Robins RW, Beer JS. 2001. Positive illusions about the self: short-term benefits and long-term costs. J Pers Soc Psychol. 80:340–352.

Roy M, Shohamy D, Wager TD. 2012. Ventromedial prefrontal-subcortical systems and the generation of affective meaning. Trends Cogn Sci. 16:147–156.

Rudebeck PH, Saunders RC, Prescott AT, Chau LS, Murray EA. 2013. Prefrontal mechanisms of behavioral flexibility, emotion regulation and value updating. Nat Neurosci. 16: 1140–1145.

Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TEJ, Johansen-Berg H, Bannister PR, De Luca M, Drobnjak I,

Flitney DE, et al. 2004. Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage. 23 (Suppl 1):S208–S219.

Smith SM, Nichols TE. 2009. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. Neuroimage. 44: 83–98.

Somerville LH, Jones RM, Ruberry EJ, Dyke JP, Glover G, Casey BJ. 2013. The medial prefrontal cortex and the emergence of self-conscious emotion in adolescence. Psychol Sci. 24: 1554–1562.

Steiner JE, Glaser D, Hawilo ME, Berridge KC. 2001. Comparative expression of hedonic impact: affective reactions to taste by human infants and other primates.. Neurosci Biobehav Rev. 25:53–74.

Tamir DI, Mitchell JJP. 2012. Disclosing information about the self is intrinsically rewarding. Proc Natl Acad Sci. 2012: 8038–8043.

Taylor SE, Brown JD. 1988. Illusion and well-being: a social psychological perspective on mental health. Psychol Bull. 103:193–210.

Taylor SE, Burklund LJ, Eisenberger NI, Lehman BJ, Hilmert CJ, Lieberman MD. 2008. Neural bases of moderation of cortisol stress responses by psychosocial resources. J Pers Soc Psychol. 95:197–211.

Wagner DD, Haxby J V, Heatherton TF. 2012. The representation of self and person knowledge in the medial prefrontal cortex. Wiley Interdiscip Rev Cogn Sci. 3:451–470.

Whalen PJ. 2007. The uncertainty of it all. Trends Cogn Sci. 11: 499–500.

Winkler AM, Ridgway GR, Webster MA, Smith SM, Nichols TE. 2014. Permutation inference for the general linear model. Neuroimage. 92:381–397.

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