The Functional Neural Architecture of Self-Reports of Affective Experience

Ajay B. Satpute, Jocelyn Shu, Jochen Weber, Mathieu Roy, and Kevin N. Ochsner

Background: The ability to self-report on affective experience is essential to both our everyday communication about emotion and our scientific understanding of it. However, the underlying cognitive and neural mechanisms for how people construct statements even as simple as "I feel bad!" remain unclear. We examined whether the neural architecture underlying the ability to make statements about affective experience is composed of distinct functional systems.

Methods: In a novel functional magnetic neuroimaging paradigm, 20 participants were shown images varying in affective intensity; they were required either to attend to and judge the affective response versus to nonaffective aspects of the stimulus and either to categorize their response into a verbal label or report on a scale that did not require verbal labeling.

Results: We found that the ability to report on affective states involves (at least) three separable systems, one for directing attention to the affective response and making attributions about it that involves the dorsomedial prefrontal cortex, one for categorizing the response into a verbal label or word that involves the ventrolateral prefrontal cortex, and one sensitive to the intensity of the affective response including the ventral anterior insula and amygdala.

Conclusions: These results suggest that unified statements about affective experience rely on integrating information from several distinct neural systems. Results are discussed in the context of how disruptions to one or another of these systems may produce unique deficits in the ability to describe affective states and the implications this may hold for clinical populations.

Key Words: Affect, emotion, fMRI, mental state attribution, neuroimaging, self-report

Across many situations and circumstances, we are called upon to answer the question, "How do I feel?" Self-reports of this sort provide the most common measure of current and enduring affective states (1) and are an integral component for undergoing several forms of psychotherapy (2,3). Given its importance (2,4), it is striking that the psychological processes used to construct such reports are so poorly understood. To be sure, many studies have examined the appraisals that trigger specific emotion self-reports and how these vary for particular kinds of people (5,6). Others have focused on the psychological consequences that using self-reports may have on affective feeling (7,8). However, the focus has rarely been on understanding how self-reports on affective states per se come into being. In part, this may be because the very ubiquity and apparent ease of providing affect reports can lead us to use them as a dependent measure rather than as the focus of research itself.

It also may be because studying the mechanisms underlying self-report using behavioral methods alone can be difficult given that they provide only indirect information about the inputs to and outputs of psychological processes. Here, neuroscience data can provide additional leverage on these processes. Accordingly, the few empirical studies on this topic have provided some insights, but they have yet to clarify which regions are most central and what functional roles they play (but see [8]). These studies ask participants to attend to and report on their affective responses and highlight activity in dorsomedial prefrontal regions implicated more generally in mental state attribution (9–11). However, activity in various other regions has also been commonly found (11–13), and in some studies, regional activations covary with the intensity of self-reports of affect (11,12,14,15), thereby raising questions about which variable (i.e., introspective self-reporting vs. affective response strength) is responsible for the findings.

To examine the processes involved in constructing self-reports of affective states, we drew on two sources. The first included psychological theories that distinguish the affective states triggered by a stimulus, which may vary continuously, from higher-level cognitive processes that can be used to attend to, semantically categorize, and verbally label these responses (16–18). On this view, affective responses in general—and self-reports of them in particular—may involve at least three types of processes: those involved in initially triggering an affective response that varies in intensity, directing attention to and becoming aware of one's resulting affective state, and the process of selecting appropriate verbal categories to describe that state. The second source was social and cognitive neuroscience research, which suggests that distinct neural regions may relate to each of these processes. The initial triggering of an affective response has been associated with a variety of regions, including the amygdala and insula (19–24). Attention to and awareness of affective states involves the attribution of mental states to the self, which has been associated with portions of dorsomedial prefrontal cortex (dmPFC) (11,18,25,26), and the categorization of various kinds of stimuli using verbal labels has been associated with portions of ventrolateral prefrontal cortex (vlPFC) (8,27–31). Taken together, these data led us to hypothesize that self-reports of affective states are constructed using the three kinds of dissociable cortical and subcortical systems described above.

To test this hypothesis, we developed the novel task depicted in Figure 1. Participants viewed images ranging from neutral/low

From the Department of Psychology (ABS), Northeastern University, Boston, Massachusetts; Department of Psychology (JS, JW, MR, KNO), Columbia University, New York, New York; and Department of Psychology and Neuroscience (MR), University of Colorado, Boulder, Colorado.

Address correspondence to Kevin Ochsner, Department of Psychology, Columbia University, 1190 Amsterdam Avenue, MC 5501, New York, NY 10025; E-mail: ochsner@psych.columbia.edu.

Received Jul 27, 2012; revised Sep 21, 2012; accepted Oct 2, 2012.

0006-3223/$36.00
http://dx.doi.org/10.1016/j.biopsych.2012.10.001

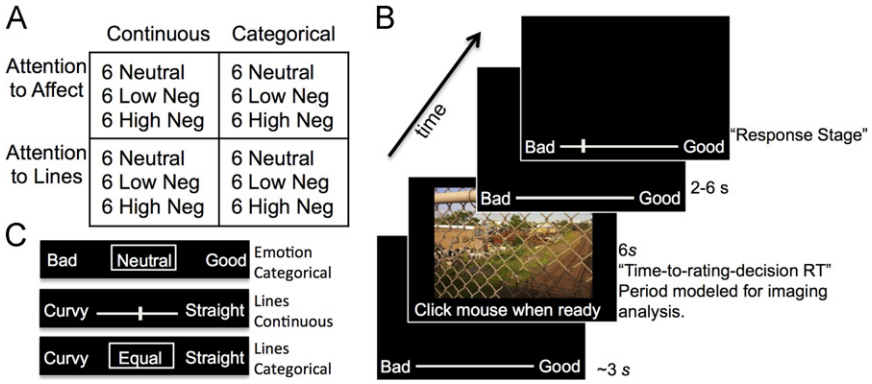

Figure 1. The experimental task: the task was designed to examine three contributions to generating a self-report of affect: attentional focus to one's affective state, which may require awareness of and attributions of one's feeling to one's self (i.e., "I feel"), categorizing that state using verbally labeled categories such as "bad" or "good" (simple labels were used here because our interest was engaging the process of categorization [16,23] rather than the particular nuances of specific affective labels per se, although in principle, the theoretical framework we use [16,23] can account for both broader and more specific labels for affective experiences), and the intensity of the affective response. Participants were shown images ranging in affective intensity, asked to attentionally focus either on their affective state (they were explicitly instructed to indicate their subjective affective response to the images and that this need not correspond to the normative characteristics of the image) or on the curviness of the lines in the image, and then indicated their current affect state using either verbal category labels or a continuous scale (on which a given point on the line did not necessarily correspond to a verbally labeled category apart from the poles). Participants determined their response while the image was in view (which corresponded to the critical period that was modeled for the imaging analyses) and then indicated their response after a jittered interval. **(A)** The experiment followed a 2 × 2 × 3 design; attentional focus and verbal categorization was manipulated across blocks and image intensity levels within blocks. **(B)** A sample trial layout for the attention to affect, continuous scale, low negative intensity image condition. The image remained on the screen for the entire 6 sec, during which participants were instructed to make a key press once they determined what their decision was for the image in terms of the previously presented scale. Regressors of interest modeled only the image-viewing phase as a 6-sec epoch, thereby keeping visuomotor components approximately identical across conditions. **(C)** Scales used for remaining conditions. For the categorical affect scale, the words "bad," "neutral," and "good" were shown without a line, and a box appeared in the decision phase that could be moved around these categories. For attention to lines (middle), the categories were "curvy," "equal," and "straight," and for the continuous scale (bottom), "equal" was replaced with a line. In between task blocks, participants completed an odd–even task for a baseline (37). The sample image in the figure is not part of the International Affective Picture System set and is shown only for display purposes. Neg, negative; RT, response time.

arousal to very negative/high arousal and judged either their affective state or the perceptual features of the image. On some trials, participants made their response using a three-category scale in which they used verbal labels to categorize their responses, and on others they used a continuously graded scale, in which points on the scale (apart from anchors) were not indicative of verbally labeled categories[1] (cf. other studies [32–35]). To be sure, participants must select or categorize their responses on both scales, but for the categorical scale, they must categorize their responses into verbal labeled categories (i.e., "good," "neutral," or "bad"), whereas on the continuous scale, they needed to select a specific point. Having participants either categorize their responses into a given verbal label or select a point on a scale maintains the same attentional focus while subtly introducing a requirement to verbally label or not. This design allowed us to dissociate the contributions to neural activity of the intensity of affective response, one's attentional focus on affect, which may involve the attribution of mental states to the self, and the categorization of an affective state into a verbal label.

To identify regions associated with attention to affect, we compared activity on attention-to-affect versus attention-to-features trials. Then, to identify regions involved in verbal categorization of affect, we compared activity on categorical versus continuous trials. Third, to rule out the possibility that regions associated with attention to or categorization of affect were covarying with affective intensity, we identified regions associated with intensity of affective response by comparing activation to very negative/high-arousal versus neutral/low-intensity images. We predicted that activity in dmPFC and vlPFC would dissociate processes associated with the attention to and categorization of affective states into a verbal label, respectively, whereas intensity should be related to activity in regions previously implicated in signaling the presence of salient affective inputs, such as the anterior insula and amygdala.

Methods and Materials

Participants

Twenty healthy, native-English-speaking, right-handed participants (aged 19–34; six male) provided informed consent following Columbia University's institutional review board guidelines. They received US$25 per hour in compensation. For two participants, only two of three scanner sessions were obtained

[1]This manipulation draws on research studies in cognitive psychology that have compared conditions that require participants to categorize their responses into discrete categories relative to various baseline conditions that range from not requiring categorization at all to reducing the demands placed on categorization. For our purposes, we took from these studies the broader point that putting affective feelings, which vary continuously and are analog, into verbal labels such as "good," "neutral," or "bad," which are discrete, involves an "analog to digital" transformation. Thus, our baseline condition of making continuous judgments follows suit in that it allows responses to remain in analog format. A second advantage of the manipulation is that it still maintains the participant's attentional focus on the same dimension (i.e., to affective states or to the lines) and hence controls for attentional focus. Critically, this manipulation is remarkably subtle, which may be necessary because more powerful manipulations of categorization (e.g., introducing multiple semantic dimensions or terms to choose from) may also influence the attentional focus.

due to scanner failure. The available scanning data and behavioral data were included in the analyses.

Stimuli

From the International Affective Picture System (36), we selected 24 neutral images, 24 low negative images, and 24 high negative images and divided them into four balanced sets that were counterbalanced across experimental conditions (see Supplement 1 for details).

Experimental Task

The experimental task was designed to differentiate neural activity associated with attending to and accessing affective mental states, categorizing those states into words, and the underlying intensity of the affective response. A $2 \times 2 \times 3$ factorial design as outlined in Figure 1 was used consisting of the factors: "attentional focus" (attend to internal affective state or attend to perceptual features of the image), "rating scale" (make ratings on a categorical or continuous scale), and "affective intensity" (neutral, low negative, and high negative images based on normative ratings, although normative ratings were correlated strongly with subjective ratings with the average correlation being $r = .74$ in the continuous condition and $r = .64$ in the categorical condition). We note that this categorization manipulation (i.e., requiring participants to categorize their responses into discrete verbal labels or allowing them to use a continuous measurement; see other resources for similar uses [32–34]) is not intended to suggest that no semantic processing is occurring in the continuous baseline but only that the relative demand or need for categorizing is greater when required to categorize using verbal labels than when required to make judgments on a continuous scale. The four combinations of attentional focus and type of rating scale were presented across 12 blocks, four per scan, block randomized. Each block contained two images of each intensity level, randomly ordered. Blocks were used to minimize switching costs between conditions and to encourage processing orientations during the image-viewing period.

At the beginning of each block an instruction cue ("How do you feel?" or "Straight or Curvy?") and corresponding scale (Figure 1) were presented for 5 sec, followed by six 14-sec trials. Trials were designed to allow separation of neural responses related to the image-viewing phase—when participants formulated their categorical or continuous judgments about the images (about either their feeling states or about the lines for the block)—from neural responses during the response stage—when they executed the key press indicating the nature of that judgment. To do this, trials began with an image shown for 6 sec with the cue, "Click mouse when ready" underneath. Participants pressed a button once they knew what response they were going to make in terms of the previously presented categorical or continuous scale. This button press provided the "time-to-rating-decision" reaction time. Upon response, the "Click" cue disappeared, but the image remained for the full 6 sec from stimulus onset to equate stimulus presentation times across conditions that were expected to differ in their time-to-rating reaction times. Then, within an 8-sec window, a decision screen was presented after a jittered interval (2, 4, or 6 sec) that consisted of the rating scale for that block (Figure 1). Participants moved the trackball to click on the appropriate part of the scale on which the bar or box disappeared. Note that this procedure ensured that, except for the imaging viewing period, the scale remained on the screen throughout the block to maintain the condition context. To acquire a separate baseline measure of neural activity unrelated to task performance, after each block of images, a block of nine odd–even judgment trials was performed for single digits presented for 2 sec each in the center of the screen along with the words "odd" and "even" on the bottom left and right of the screen (37). Participants pressed the left and right buttons for odd and even numbers, respectively. Participants were provided with practice to orient them to the task (Supplement 1) and were explicitly instructed to indicate their own subjective affective responses to the images.

Apparatus

Scanning was conducted on a GE (Fairfield, Connecticut) TwinSpeed 1.5-T scanner equipped with an eight-channel head coil. Functional scans were obtained using a spiral in/out pulse sequence, and structural scans were obtained using an spoiled gradient recoil sequence (see Supplement 1 for more details and stimulus setup). Stimuli were projected on a screen visible in a mirror attached to the head coil. Responses were made with a scanner compatible trackball.

Data Analysis

Functional images were preprocessed in SPM5, and statistical models were implemented using a combination of BrainVoyager (Brain Innovation, Maastricht, The Netherlands) and NeuroElf (www.neuroelf.com) software packages. Images were coregistered, motion corrected, normalized (MNI-ICBM152 template), resliced (3 mm^3 voxels), and smoothed (6-mm full width at half maximum). First-level models included separate image-viewing regressors for each of the 12 conditions of the $2 \times 2 \times 3$ design, which were modeled as 6-sec epochs. Also included were a single nuisance regressor controlling for motor responses in the decision logging phase (from decision screen onset to response across conditions) and a high-pass filter (220 sec cutoff). Regressors were convolved with the canonical hemodynamic response function. Robust regressions were performed at the first level to reduce the influence of outliers in estimating model fit.

For second-level analyses, subjects were modeled as a random variable, and an AlphaSim MonteCarlo simulation as implemented in the Analysis for Functional NeuroImages (AFNI; Bethesda, Maryland) software (smoothing kernel estimated at 7 mm from the data) was used to select a combined height ($p < .002$) and extent ($k = 20$) threshold to identify clusters that resulted in a whole-brain family-wise error corrected threshold of $p < .05$ (11,38–42). Using this corrected threshold, an omnibus F test was used to identify clusters that showed significant variability due to the experimental conditions. We then extracted and averaged beta values for voxels within each cluster to produce one value per condition per subject per cluster. We compared clusters in our primary regions of interest (ROIs), the dmPFC and vlPFC, using a 2 (region) $\times 2 \times 2 \times 3$ repeated-measures analysis of variance (ANOVA) model to examine our hypothesized interactions between these two areas. For remaining clusters, we used $2 \times 2 \times 3$ repeated-measures ANOVAs to identify how each factor contributed to variability in these regions. We further controlled for differences in reaction time across the attention to emotion and attention to lines conditions by including it as a covariate in an analysis of covariance model. If reaction time reduced the significance of the result, the reduced statistical values were reported (as indicated in Table 1; see footnote c), or the cluster was removed from further analysis if it reduced significance below threshold. We performed outlier correction by calculating the Mahalanobis distance and excluding values more than 3 SD from the mean.

Table 1. Neural Regions Responsive to the Attention, Categorization, and Intensity of Affect

Area	BA	Direction	MNI			*k*	Effect Sizes		
			x	y	z		Attention	Categorization	Intensity
Attention to Affect vs. Attention to Lines									
dmPFC	9	Affect > lines	9	63	30	292	**.61**[a]	.12	.02
FP	10/11	Affect > lines	18	57	−9	38	**.62**[a]	.09	.11
SFG	9	Affect > lines	−15	54	36	158	**.53**[a]	.12	.02
MFG	9	Affect > lines	−33	51	33	29	**.35**[b]	.03	.09
Mid-CING	23/24	Affect > lines	−9	−9	27	185	**.49**[b]	.11	.09
STS	22	Affect > lines	−51	−15	−15	50	**.33**[b,c]	.16	.13
TPJ	39	Affect > lines	−51	−51	27	36	**.52**[a]	.01	.04
Cerebellum		Affect > lines	21	−78	−36	56	**.50**[b,c]	.01	.01
Categorization into a Verbal Label vs. Selection on a Continuously Graded Scale									
vlPFC	44	Verbal label > continuous	60	18	6	26	.03	**.45**[a]	.15
SPL	7	Verbal label > continuous	12	−72	57	49	.19[c]	**.38**[b]	.2
Affective Intensity: High, Low, Neutral									
Precentral gyrus	4/6	High negative > neutral	57	0	33	28	.26	.01	**.35**[a]
LING		High negative > neutral	12	−51	−6	29	.03	.09	**.34**[a]
Fusiform		High negative > neutral	−36	−54	−18	34	.05	.06	**.53**[a]
ITG	19	High negative > neutral	−48	−69	−6	457	0	.02	**.55**[a]
SLEA		High negative > neutral	−6	3	−15	40	.21[c]	.01	**.44**[a]
Attention to Affect and Affective Intensity									
Precuneus	23/30/31	High negative > neutral, affect > lines	−3	−60	24	761	**.40**[b,c]	.2	**.35**[a]
TPJ	39	High Negative > neutral, affect > lines	57	−60	15	125	**.46**[b,c]	.02	**.57**[a]
Putamen		High negative > neutral	−24	12	−12	112	**.41**[b,c]	.05	**.43**[a]
Categorization into a Verbal Label and Affective Intensity									
ITG		Categorical > continuous, high negative > neutral	54	−66	−6	284	.03	**.32**[b]	**.73**[a]

The table illustrates that several neural areas commonly involved in affect and emotion show sensitivity to specific components of reporting on affective experience including attentional focus on affect (involving the attribution of mental states to the self, e.g. "I feel . . ."), verbal categorization (involving the act of categorizing affective states using labels such as "bad," "Neutral," or "Good"; simple labels were used here because our interest was engaging the act of categorization rather than the particular nuances of specific labels per se), and the intensity of affective experience (involving the strength of the affective response ranging and providing information). Clusters were observed from an omnibus *F* test (see Methods and Materials). The last three columns refer to effect sizes of the three manipulations: attentional focus, categorization using verbal labels, and affective intensity. Effect sizes are included here to illustrate relative differences in effects across the manipulations, but absolute effect sizes should be interpreted with caution. Because the repeated-measures design involves computing distinct error variability for each factor, the magnitude of effect size required to obtain significance varied slightly depending on the factor. Bolded effect sizes for each main effect are significant. BAs are putative.

BA, Brodmann's area; CING, cingulate cortex; dmPFC, dorsomedial prefrontal cortex; FP, frontopolar cortex; INS, insular cortex; IPS, intraparietal sulcus; ITG, inferior temporal gyrus; LING, lingual gyrus; MFG, middle frontal gyrus; MNI, Montreal Neurological Institute; MTG, middle temporal gyrus; SFG, superior frontal gyrus; SLEA, sublenticular extended amygdala; SPL, superior parietal lobule; STS, superior temporal sulcus; TPJ, temporoparietal junction; vlPFC, ventrolateral prefrontal cortex.

[a] $p < .01$

[b] $p < .001$.

[c] Results that were reduced by including reaction time as a covariate.

Results

Behavioral Results

Time-to-Rating-Decision Reaction Times. Time-to-rating-decision reaction times were analyzed using a $2 \times 2 \times 3$ repeated-measures ANOVA. A significant main effect of attentional focus was found, indicating that judging lines (in seconds, mean reaction time = 2.96) took longer than judging affect [mean reaction time = 2.75 sec; $F(1,17) = 18.36$, $p < .01$]. This suggests that making line judgments may have been more difficult than making affect judgments. Hence, we controlled for reaction time in the neuroimaging results that follow by including it as a covariate in an analysis of covariance model. Results are reported for the model with the covariate when reaction time influenced the results and without the covariate otherwise. Critically, reaction time did not influence activity in our regions of a priori interest for responding to attending to affect, the dmPFC, nor for categorizing affective experience into verbal labels, the vlPFC. No main effects of rating scale ($p > .4$) or intensity ($p > .3$) on reaction time were found. Finally, a significant interaction was found between attentional focus and intensity [$F(2,34) = 4.50$, $p < .05$]. However, post hoc tests examining reaction time differences in intensity within attentional focus conditions were not significant ($ps > .3$; Figure S1 in Supplement 1).

Self-Reports of Affective States. Self-reports of negative affect were analyzed for attention to affect trials. Subjective ratings correlated strongly with normative stimulus ratings ($rs > .6$; see Methods and Materials). High negative images produced higher ratings of negativity than low negative, and low negative than neutral, in both the continuous [$t(19) = 7.40$, $p < .001$; $t(19) = 5.38$, $p < .001$, respectively] and categorical [$t(19) = 7.17$, $p < .001$; $t(19) = 7.00$, $p < .001$, respectively] rating

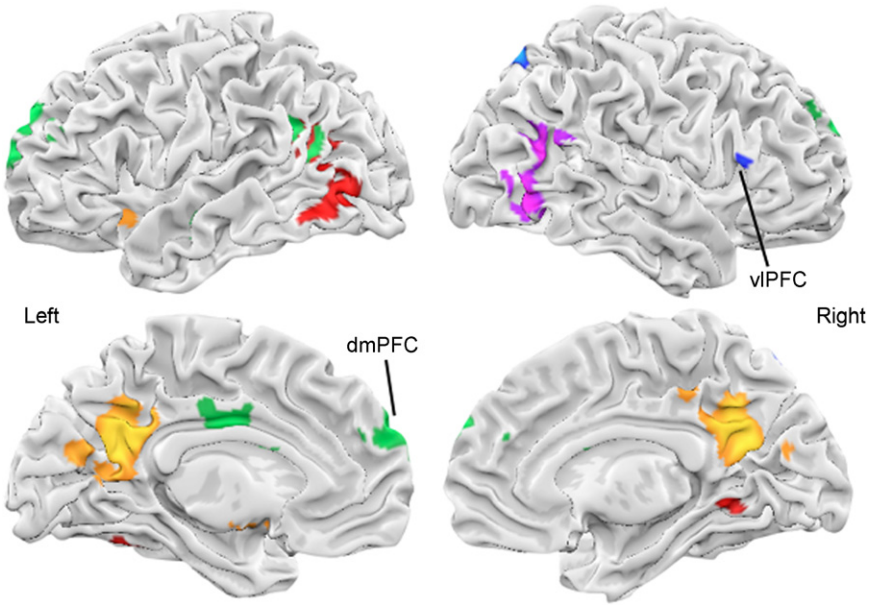

Figure 2. The overall set of neural regions that were responsive to conditions in the experimental task. An omnibus *F* test was used to identify clusters ($p < .05$, family-wise error corrected). Activity in voxels was then averaged within each cluster to produce one value for each of the 12 conditions, per cluster, per subject. Using repeated-measures analysis of variance models, we examined how each of three factors contributed to activity in these regions during the image-viewing period (see task design): attention to and awareness of one's affective state, categorization of that state using verbal labels such as "bad" or "good" (simple labels were used here because our interest was engaging the act of categorization rather than the particular nuances of specific affective labels per se), and the intensity of the affective response. The color codes used in the image were assigned based on whether effects contributed significantly to the clusters (based on the values listed in Table 1). Green, attention to affect; blue, categorization into verbal labels; red, affective intensity; yellow, attention to affect and affective intensity; purple, categorization into verbal labels and affective intensity. DMPFC, dorsomedial prefrontal cortex; VLPFC, ventrolateral prefrontal cortex.

scale conditions. Although responses on the two scales were not directly comparable, similar increases were observed regardless of the scale used.

Neuroimaging Results

For all imaging analyses reported here, activity during only the image-viewing period of the trials was compared across conditions (Figure 1B). The imaging results are summarized in Table 1 and Figure 2, which shows how specific factors contributed to variability in neural activity for clusters that were responsive in the task.

Attention to Affect Versus Perceptual Features. Activity in dmPFC was greater when attending to affect but was not influenced by verbal categorization or affective intensity (Table 1; also see ROI analysis that follows). Other regions engaged included mid-cingulate cortex, temporoparietal junction, and superior temporal cortex (Table 1). Areas showing greater activity when attending to affect than to perceptual features and to increasing affective intensity included the right temporoparietal junction and precuneus. These results are consistent with several prior reports showing greater activity in similar areas when attending to affective states (9–11), and more broadly, when engaging in mental state attribution (26).

Categorization into a Verbal Label Versus Selection on a Continuously Graded Scale. Activity in right vlPFC, superior parietal lobule, and inferior temporal cortex was greater during verbal categorization than when making judgments on a continuously graded scale (see Table 1; see also ROI analysis). Given that a main goal of this article was to test the hypothesis that vlPFC is involved in the act of categorizing affective states, we specifically tested for and found that the effect in right vlPFC was significant when categorizing affect [$t(19) = 3.66$, $p < .002$]. It was not significantly greater for categorizing lines [upon removal of one outlier; $t(18) = 1.61$, $p = .13$], and the interaction was marginally significant [$F(1,18) = 3.79$, $p = .067$]. This pattern of results is generally consistent with prior studies examining verbal labeling of affective stimuli in general (8,28,29).

High Versus Low Arousal Stimuli. Several neural regions were sensitive to increasing negative affective intensity, some of which were also responsive to attention to affect (Table 1). These included the left ventral anterior insula, portions of temporal cortex, and subcortical regions including the thalamus, brainstem (extending from the lingual gyrus), and a cluster extending from

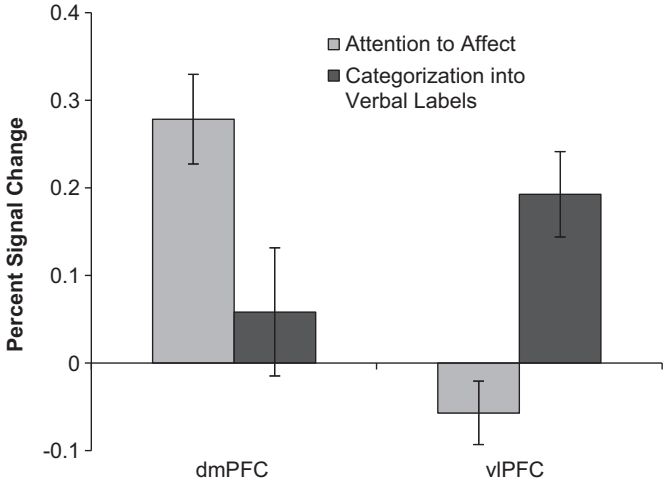

Figure 3. The figure illustrates the effects of attention to affect and categorization of affect into verbal labels on neural activity in dorsomedial prefrontal cortex (DMPFC) and right ventrolateral prefrontal cortex (VLPFC) regions of interest. Activity during the image-viewing phase was greater in the DMPFC when attending to affective experience, but was not influenced by verbally categorizing (see Supplementary Materials for means separated by conditions). In contrast, activity during the image-viewing phase in right VLPFC was greater when verbally categorizing, for both categorizing affective experience into verbal labels or when categorizing the lines in the image (i.e. in a domain general manner), but showed no effect of attentional focus. These results suggest that attentional focus on and the categorization of affective states using verbal labels are distinct psychological processes that are associated with separable neural systems. Results are presented in summary format here, but see Figure S2 in Supplement 1 for the marginal means for each condition.

the sublenticular extended amygdala and into the amygdala (43). No areas showed inverse relationships with affective intensity.

ROI Analyses: Dissociating Processes Underlying Self-Reports of Affective Experience. To test the hypothesis that an attentional focus on and the verbal categorization of affective states are dissociable components for the behavior of reporting on those states, we tested for hypothesized interactions between dmPFC and vlPFC within a full 2 (region) × 2 × 2 × 3 ANOVA model (which controlled for intensity). As illustrated in Figure 3, a significant 2 (region) × 2 (attentional focus) interaction [$F(1,19) = 26.50$, $p < .0001$] indicated that the effect of attentional focus depended on the region involved: dmPFC showed greater activity when attending to affect [$t(19) = 5.45$, $p < .0001$], whereas vlPFC did not [$t(19) = -.78$, $p = .45$]. Furthermore, a significant 2 (region) × 2 (categorization) interaction [$F(1,19) = 7.71$, $p = .012$] showed that the effect of categorizing also depended on the region: vlPFC showed greater activity during categorizing [$t(19) = 3.95$, $p = .001$], whereas dmPFC did not [$t(19) = 1.61$, $p = .12$]. These interactions illustrate a double dissociation of functional activity between dmPFC and right vlPFC. That is, two experimental manipulations, attention to affect and the effect of categorizing into verbal labels, were shown to have different effects on two dependent variables, activity in dmPFC and activity in vlPFC (Figure 3).

Discussion

We began with the question of what psychological and neural mechanisms underlie our ability to report on our affective states introspectively. We reasoned that making statements even as simple as "I feel good" or "I feel bad" that communicate a unified affective sentiment may involve the engagement of distinct neural systems (23). Using a novel task design, we found that attending to affect engaged the dmPFC, categorization of that state into a verbal label engaged the vlPFC, and the intensity of affective response engaged the amygdala and insula.

Implications for Understanding Self-Reports of Affective Experience

Together, these results support the view that describing affective experiences with language is a constructive act (44) that may depend on three types of separable systems. The first system relying on the dmPFC supports directing attention to affective states and, along with activation in the temporoparietal junction and precuneus, may more broadly reflect the ability to make mental state attributions (11,25,26). In support of this, damage to the dmPFC (albeit not isolated to the dmPFC in these studies) appears to affect the awareness of having, but not the production of, affective responses (45,46). A second system relying on the vlPFC is involved in placing affective experience into available semantic categories. Previous studies in humans and nonhuman primates have found the vlPFC to be a critical region for both category learning and retrieval[2] (30,31). Although many of these studies have focused on left vlPFC for general semantic categorization, right vlPFC has been associated with categorization along perceptual (47) and affective dimensions (8,28,29). These results extend the role of right vlPFC established in previous studies, which have focused on labeling qualities of affective stimuli, to include the labeling of qualities of subjective internal affective experiences. This region may draw on temporal cortical regions (48) to connect affective responses with a body of semantic knowledge.

Finally, a third system relying on the amygdala and anterior ventral insula, among other areas, is involved in producing affective responses (23,49). This system is engaged by stimuli of increasing affective intensity. For some regions, this may be regardless of reflective awareness (19); indeed, damage to the amygdala or the insula does not seem to impair the ability to report on affective states (50,51), although it may dampen the subjective intensity of affective responses (51–53).

In summary, the results suggest that putting feelings into expressible statements such as "I feel bad!" involves three components: self-reflective attention to and attribution of affective states (i.e., "I feel") (26,54), categorization of those state using verbal labels (i.e., "good," "bad," etc.), and the intensity of the affective response.

Implications for Theories of Affect and Emotion

These data have implications for theories of affect and emotion in both neuroscience and psychology. Neuroscience research has produced a wealth of data describing how neural regions are associated with responses to various kinds of affective stimuli (20). However, such experiments have not clearly identified the specific processes supporting our ability to reflect on and report our internal affective experiences. On the psychology side, theories have suggested multiple processes are involved in constructing an affective experience (44,55) but have seldom tested these notions directly (56). By using functional magnetic resonance imaging to test predictions made by psychological theories of emotion, our study joins a growing number demonstrating the utility of integrating neuroscience data with emotion theory to unpack the specific functions performed by systems generally implicated in emotion (8,22,57). In doing so, we may move toward more comprehensive theories of the neural systems, supporting a wide range of affective phenomena.

Although the present study examined the use of general affective labels (i.e., judging whether one feels "good" or "bad"), in principle these findings may apply to the use of more specific emotion words such as "happy" or "angry." For example, one of the psychological theories that motivated this study posits that continuously graded affective responses only become emotions when one semantically categorizes them (16). Future work could test whether the use of discrete emotion words to categorize affective states would rely on the same vlPFC system identified here. Indeed, previous research that has examined the labeling of facial expression stimuli suggests that this may be the case (8,29).

Implications for the Study of Healthy and Clinical Differences in Affective Experience

Our findings raise novel questions for understanding individual variability in affective experience. In children, the ability to identify and describe one's own emotions emerges around 26 months of age and forms an integral part in how parents shape and train children's affective reactions and their regulation (58). This research opens the door to asking which processes develop first and with what fidelity—being able to introspect on affective states, which may rely on ascribing mental states to the self, or

[2]Notably, the vlPFC activations found here and in previous studies of verbal labeling of affect are inferior to the dorsolateral regions responsive to increasing working memory demand (meta-analysis by Owen *et al.* [72]).

learning categories/labels for one's feelings and then using them to guide introspection in constructing emotional experiences.

Deficits in the ability to introspect on affect and emotion have been associated with greater symptom severity, poorer treatment outcomes, and ineffective treatment by psychotherapy across several mental health disorders, as measured by alexithymia. These deficits are present in a variety of conditions, including autism spectrum disorder (59), schizophrenia (60), and major depressive disorder (61,62). Intriguingly, the dorsomedial and ventrolateral prefrontal regions we observed to relate to different components of reporting on affective states also show distinct functional and structural relationships to these conditions (63–71). Thus, interruptions in any of the component processes underlying self-reports of affect may ultimately lead to disturbances in the experience and regulation of emotion and may do so in unique ways depending on which systems are compromised in these populations. Future studies investigating these abilities in clinical populations may lead to new directions that either target compromised processes or leverage more intact processes to promote better self-awareness of emotional states.

Conclusions

In psychological and neural studies, the ability to report on our affective states is often taken for granted as an output measure for emotion. However, recent theories of emotion suggest that our ability to make such reports also relies on processes that are integral to both constructing an affective experience and regulating it (16). This study brought together diverse themes from emotion theory, the cognitive psychology of categorization, and findings in affective neuroscience , to help unpack the functional architecture underlying the ability to produce self-reports of internal affective experiences. In doing so, new questions can be raised regarding how component processes of this ability play roles in various affective phenomena. Ongoing research may open up novel avenues for understanding how the ability to self-report on emotion varies across development and contributes to mental health issues.

This research was supported by National Institutes of Health Grant No. 1 R01 MH076137-01 A1 awarded to Kevin Ochsner.

We thank Andrew Kogan for assistance in magnetic resonance imaging data acquisition.

The authors report no biomedical financial interests or potential conflicts of interest.

Supplementary material cited in this article is available online.

1. Beck A, Ward CH, Mendelson M, Mock J, Erbaugh J (1961): An inventory for measuring depression. *Arch Gen Psychiatry* 4:561–571.
2. Sifneos PE (1973): The prevalence of "alexithymic" characteristics in psychosomatic patients. *Psychother Psychosom* 22:255–262.
3. Taylor GJ, Bagby RM, Parker JDA (1999): *Disorders of Affect Regulation: Alexithymia in Medical and Psychiatric Illness*. Cambridge, UK: Cambridge University Press.
4. Zaki J, Bolger N, Ochsner K (2008): It takes two: The interpersonal nature of empathic accuracy. *Psychol Sci* 19:399–404.
5. Scherer KR, Schorr A, Johnstone T (2001): *Appraisal Processes in Emotion: Theory, Methods, Research*. Oxford, England: Oxford University Press.
6. Lewis M, Haviland-Jones JM, Barrett LF (2010): *Handbook of Emotions, 3rd ed*. New York: Guilford Press.
7. Pennebaker JW (1997): Writing about emotional experiences as a therapeutic process. *Psychological Science* 8:162–166.
8. Lieberman MD, Eisenberger NI, Crockett MJ, Tom SM, Pfeifer JH, Way BM (2007): Putting feelings into words. *Psychological Science* 18:421-421.
9. Gusnard DA, Akbudak E, Shulman GL, Raichle ME (2001): Medial prefrontal cortex and self-referential mental activity: relation to a default mode of brain function. *Proc Natl Acad Sci U S A* 98: 4259–4264.
10. Lane RD, Fink GR, Chau PM, Dolan RJ (1997): Neural activation during selective attention to subjective emotional responses. *Neuroreport* 8: 3969–3972.
11. Ochsner KN, Knierim K, Ludlow DH, Hanelin J, Ramachandran T, Glover G, *et al*. (2004): Reflecting upon feelings: An fMRI study of neural systems supporting the attribution of emotion to self and other. *J Cogn Neurosci* 16:1746–1772.
12. Goldin PR, Hutcherson CA, Ochsner KN, Glover GH, Gabrieli JD, Gross JJ (2005): The neural bases of amusement and sadness: A comparison of block contrast and subject-specific emotion intensity regression approaches. *Neuroimage* 27:26–36.
13. Hutcherson CA, Goldin PR, Ochsner KN, Gabrieli JD, Barrett LF, Gross JJ (2005): Attention and emotion: Does rating emotion alter neural responses to amusing and sad films? *Neuroimage* 27:656–668.
14. Phan KL, Taylor SF, Welsh RC, Ho SH, Britton JC, Liberzon I (2004): Neural correlates of individual ratings of emotional salience: A trial-related fMRI study. *Neuroimage* 21:768–780.
15. Taylor SF, Phan KL, Decker LR, Liberzon I (2003): Subjective rating of emotionally salient stimuli modulates neural activity. *Neuroimage* 18: 650–659.
16. Barrett LF (2006): Solving the emotion paradox: Categorization and the experience of emotion. *Pers Soc Psychol Rev* 10:20–46.
17. Ochsner KN, Barrett LF (2001): A multiprocess perspective on the neuroscience of emotion. In: Mayne TJ, Bonanno GA, editors *Emotions: Currrent Issues and Future Directions*. New York, NY: Guilford Press, 38–81.
18. Olsson A, Ochsner KN (2008): The role of social cognition in emotion. *Trends Cogn Sci* 12:65–71.
19. Whalen PJ (1998): Fear, vigilance, and ambiguity: Initial neuroimaging studies of the human amygdala. *Curr Direct Psychol Sci* 7:177–188.
20. Phan KL, Wager T, Taylor SF, Liberzon I (2002): Functional neuroanatomy of emotion: A meta-analysis of emotion activation studies in PET and fMRI. *Neuroimage* 16:331–348.
21. Craig AD (2002): How do you feel? Interoception: The sense of the physiological condition of the body. *Nat Rev Neurosci* 3:655–666.
22. Wager TD, Barrett LF, Bliss-Moreau E, Lindquist K, Duncan S, Kober H, *et al*. (2008): The neuroimaging of emotion. In: Lewis M, Haviland-Jones LF, Barrett LF, editors *Handbook of Emotions, 2nd ed*. New York: Guilford, 513–530.
23. Barrett LF, Mesquita B, Ochsner KN, Gross JJ (2007): The experience of emotion. *Annu Rev Psychol* 58:373–403.
24. Phelps EA, LeDoux JE (2005): Contributions of the amygdala to emotion processing: From animal models to human behavior. *Neuron* 48:175–187.
25. Gallagher HL, Frith CD (2003): Functional imaging of "theory of mind". *Trends Cogn Sci* 7:77–83.
26. Amodio DM, Frith CD (2006): Meeting of the minds: The medial frontal cortex and social cognition. *Nat Rev Neurosci* 7:268–277.
27. Lieberman MD (2007): Social cognitive neuroscience: A review of core processes. *Annu Rev Psychol* 58:259–289.
28. Lieberman MD (2011): Why symbolic processing of affect can disrupt negative affect: Social cognitive and affective neuroscience investigations. In: Todorov AB, Fiske ST, Prentice DA, editors *Social Neuroscience: Toward Understanding the Underpinnings of the Social Mind*. New York: Oxford University Press, 188–209.
29. Hariri AR, Bookheimer SY, Mazziotta JC (2000): Modulating emotional responses: Effects of a neocortical network on the limbic system. *Neuroreport* 11:43–48.
30. Freedman DJ, Riesenhuber M, Poggio T, Miller EK (2001): Categorical representation of visual stimuli in the primate prefrontal cortex. *Science* 291:312–316.
31. Passingham RE, Toni I, Rushworth MFS (2000): Specialisation within the prefrontal cortex: The ventral prefrontal cortex and associative learning. *Exp Brain Res* 133:103–113.
32. Kay P, Kempton W (1984): What is the Sapir-Whorf hypothesis? *Am Anthropol* 86:65–79.

33. Harnad S, editor (1990). *Categorical Perception: The Groundwork of Cognition*. New York: Cambridge University Press.
34. Goldstone R (1994): Influences of categorization on perceptual discrimination. *J Exp Psychol Gen* 123:178–200.
35. Master A, Markman EM, Dweck CS (2012): Thinking in categories or along a continuum: Consequences for children's social judgments. *Child Dev* 83:1145–1163.
36. Lang PJ, Bradley MM, Cuthbert BN (1997): *International Affective Picture System (IAPS): Technical Manual and Affective Ratings*. Gainesville, FL: Center for the Study of Emotion and Attention.
37. Stark CE, Squire LR (2001): When zero is not zero: The problem of ambiguous baseline conditions in fMRI. *Proc Natl Acad Sci U S A* 98: 12760–12766.
38. Bennett CM, Wolford GL, Miller MB (2009): The principled control of false positives in neuroimaging. *Soc Cogn Affect Neurosci* 4:417–422.
39. Forman SD, Cohen JD, Fitzgerald M, Eddy WF, Mintun MA, Noll DC (1995): Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. *Magn Reson Med* 33:636–647.
40. Kober H, Mende-Siedlecki P, Kross EF, Weber J, Mischel W, Hart CL, *et al*. (2010): Prefrontal-striatal pathway underlies cognitive regulation of craving. *Proc Natl Acad Sci U S A* 107:14811–14816.
41. Somerville LH, Kelley WM, Heatherton TF (2010): Self-esteem modulates medial prefrontal cortical responses to evaluative social feedback. *Cereb Cortex* 20:3005–3013.
42. Zaki J, Weber J, Bolger N, Ochsner K (2009): The neural bases of empathic accuracy. *Proc Natl Acad Sci U S A* 106:11382–11387.
43. Liberzon I, Phan KL, Decker LR, Taylor SF (2003): Extended amygdala and emotional salience: A PET activation study of positive and negative affect. *Neuropsychopharmacology* 28:726–733.
44. Barrett LF (2006): Are emotions natural kinds? *Perspect Psychol Sci* 1; 28–28.
45. Hornak J, Bramham J, Rolls ET, Morris RG, O'Doherty J, Bullock PR (2003): Changes in emotion after circumscribed surgical lesions of the orbitofrontal and cingulate cortices. *Brain* 126:1691–1712.
46. Sturm VE, Rosen HJ, Allison S, Miller BL, Levenson RW (2006): Self-conscious emotion deficits in frontotemporal lobar degeneration. *Brain* 129:2508–2516.
47. Dobbins IG, Wagner AD (2005): Domain-general and domain-sensitive prefrontal mechanisms for recollecting events and detecting novelty. *Cereb Cortex* 15:1768–1778.
48. Badre D, Wagner AD (2007): Left ventrolateral prefrontal cortex and the cognitive control of memory. *Neuropsychologia* 45:2883–2901.
49. Craig AD (2009): How do you feel—now? The anterior insula and human awareness. *Nat Rev Neurosci* 10:59–70.
50. Anderson AK, Phelps EA (2002): Is the human amygdala critical for the subjective experience of emotion? Evidence of intact dispositional affect in patients with amygdala lesions. *J Cogn Neurosci* 14:709–720.
51. Berntson GG, Norman GJ, Bechara A, Bruss J, Tranel D, Cacioppo JT (2011): The insula and evaluative processes. *Psychol Sci* 22:80–86.
52. Glascher J, Adolphs R (2003): Processing of the arousal of subliminal and supraliminal emotional stimuli by the human amygdala. *J Neurosci* 23:10274–10282.
53. Ochsner KN, Phelps E (2007): Emerging perspectives on emotion-cognition interactions. *Trends Cogn Sci* 11:317–318.
54. 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.
55. Clore GL, Ortony A (2008): Appraisal theories: How cognition shapes affect into emotion. In: Lewis M, Haviland-Jones JM, Barrett LF, editors. *Handbook of Emotions, 3rd ed*. New York: Guilford Press, 628–642.
56. Robinson MD, Clore GL (2002): Episodic and semantic knowledge in emotional self-report: Evidence for two judgment processes. *J Pers Soc Psychol* 83:198–215.
57. Ochsner KN, Gross JJ (2005): The cognitive control of emotion. *Trends Cogn Sci* 9:242–249.
58. Bretherton I, Fritz J, Zahnwaxler C, Ridgeway D (1986): Learning to talk about emotions—a functionalist perspective. *Child Dev* 57: 529–548.
59. Hill E, Berthoz S, Frith U (2004): Cognitive processing of own emotions in individuals with Autism spectrum disorder and in their relatives. *J Autism Dev Disord* 34:229–235.
60. van 't Wout M, Aleman A, Bermond B, Kahn RS (2007): No words for feelings: Alexithymia in schizophrenia patients and first-degree relatives. *Compr Psychiatry* 48:27–33.
61. Honkalampi K, Hintikka J, Tanskanen A, Lehtonen J, Viinamaki H (2000): Depression is strongly associated with alexithymia in the general population. *J Psychosom Res* 48:99–104.
62. Lipsanen T, Saarijarvi S, Lauerma H (2004): Exploring the relations between depression, somatization, dissociation and alexithymia—overlapping or independent constructs? *Psychopathology* 37:200–206.
63. Di Martino A, Ross K, Uddin LQ, Sklar AB, Castellanos FX, Milham MP (2009): Functional brain correlates of social and nonsocial processes in autism spectrum disorders: An activation likelihood estimation meta-analysis. *Biol Psychiatry* 65:63–74.
64. Happe F, Ehlers S, Fletcher P, Frith U, Johansson M, Gillberg C, *et al*. (1996): "Theory of mind" in the brain. Evidence from a PET scan study of Asperger syndrome. *Neuroreport* 8:197–201.
65. Duerden EG, Mak-Fan KM, Taylor MJ, Roberts SW (2012): Regional differences in grey and white matter in children and adults with autism spectrum disorders: An activation likelihood estimate (ALE) meta-analysis. *Autism Res* 5:49–66.
66. Taylor SF, Kang J, Brege IS, Tso IF, Hosanagar A, Johnson TD (2012): Meta-analysis of functional neuroimaging studies of emotion perception and experience in schizophrenia. *Biol Psychiatry* 71:136–145.
67. Bora E, Fornito A, Radua J, Walterfang M, Seal M, Wood SJ, *et al*. (2011): Neuroanatomical abnormalities in schizophrenia: A multimodal voxelwise meta-analysis and meta-regression analysis. *Schizophr Res* 127:46–57.
68. Ochsner KN (2008): The social-emotional processing stream: Five core constructs and their translational potential for schizophrenia and beyond. *Biol Psychiatry* 64:48–61.
69. Anand A, Li Y, Wang Y, Wu J, Gao S, Bukhari L, *et al*. (2005): Activity and connectivity of brain mood regulating circuit in depression: a functional magnetic resonance study. *Biol Psychiatry* 57:1079–1088.
70. Beauregard M, Leroux JM, Bergman S, Arzoumanian Y, Beaudoin G, Bourgouin P, *et al*. (1998): The functional neuroanatomy of major depression: An fMRI study using an emotional activation paradigm. *Neuroreport* 9; 3253–3253.
71. Lawrence NS, Williams AM, Surguladze S, Giampietro V, Brammer MJ, Andrew C, *et al*. (2004): Subcortical and ventral prefrontal cortical neural responses to facial expressions distinguish patients with bipolar disorder and major depression. *Biol Psychiatry* 55:578–587.
72. Owen AM, McMillan KM, Laird AR, Bullmore E (2005): N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Hum Brain Mapp* 25:46–59.