Behavioral/Cognitive

Human Amygdala Tracks a Feature-Based Valence Signal Embedded within the Facial Expression of Surprise

M. Justin Kim,[1,2] Alison M. Mattek,[1] Randi H. Bennett,[3] Kimberly M. Solomon,[1] Jin Shin,[1] and Paul J. Whalen[1]

[1]Department of Psychological and Brain Sciences, Dartmouth College, Hanover, New Hampshire 03755, [2]Department of Psychology and Neuroscience, Duke University, Durham, North Carolina 27708, and [3]Department of Psychology, Fordham University, Bronx, New York 10458

Human amygdala function has been traditionally associated with processing the affective valence (negative vs positive) of an emotionally charged event, especially those that signal fear or threat. However, this account of human amygdala function can be explained by alternative views, which posit that the amygdala might be tuned to either (1) general emotional arousal (activation vs deactivation) or (2) specific emotion categories (fear vs happy). Delineating the pure effects of valence independent of arousal or emotion category is a challenging task, given that these variables naturally covary under many circumstances. To circumvent this issue and test the sensitivity of the human amygdala to valence values specifically, we measured the dimension of valence within the single facial expression category of surprise. Given the inherent valence ambiguity of this category, we show that surprised expression exemplars are attributed valence and arousal values that are uniquely and naturally uncorrelated. We then present fMRI data from both sexes, showing that the amygdala tracks these consensus valence values. Finally, we provide evidence that these valence values are linked to specific visual features of the mouth region, isolating the signal by which the amygdala detects this valence information.

Key words: ambiguity; amygdala; emotion; face; surprise; valence

Significance Statement

There is an open question as to whether human amygdala function tracks the valence value of cues in the environment, as opposed to either a more general emotional arousal value or a more specific emotion category distinction. Here, we demonstrate the utility of surprised facial expressions because exemplars within this emotion category take on valence values spanning the dimension of bipolar valence (positive to negative) at a consistent level of emotional arousal. Functional neuroimaging data showed that amygdala responses tracked the valence of surprised facial expressions, unconfounded by arousal. Furthermore, a machine learning classifier identified particular visual features of the mouth region that predicted this valence effect, isolating the specific visual signal that might be driving this neural valence response.

Introduction

Understanding emotional signals conveyed by the facial expressions of others is key to successful social interaction. One of the most basic components of such emotional signals is affective valence, the degree of positivity-negativity or pleasantness-unpleasantness (Russell, 1980). Typically, facial expressions of emotion convey a clear signal with regards to their valence: for example, a person with a happy expression could be interpreted to be experiencing a positive emotion. One notable exception is surprise, a facial expression that could be perceived as either positive or negative; in other words, surprised faces are characterized as being ambiguous with respect to valence (Tomkins and McCarter, 1964; Mattek et al., 2017).

fMRI studies have traditionally documented that the human amygdala is highly responsive to negative facial expressions, including fear (Breiter et al., 1996). Meanwhile, other fMRI studies have shown that the amygdala is responsive to positive as well as negative facial expressions (Fitzgerald et al., 2006), contributing to a view that the amygdala may be generally sensitive to emotional significance (Anderson and Phelps, 2001) or socially salient information (Adolphs, 2010). Indeed, a number of fMRI studies have suggested that the amygdala might be better understood as tracking the arousal, rather than the valence value, of an emotional stimulus (Anderson et al., 2003; e.g., Wilson-Mendenhall et al., 2013). The arousal account of amygdala func-

Received May 19, 2017; revised Aug. 22, 2017; accepted Aug. 25, 2017.

Author contributions: M.J.K. and P.J.W. designed research; M.J.K., R.H.B., K.M.S., and J.S. performed research; A.M.M., K.M.S., and J.S. contributed unpublished reagents/analytic tools; M.J.K., A.M.M., R.H.B., K.M.S., J.S., and P.J.W. analyzed data; M.J.K., A.M.M., R.H.B., and P.J.W. wrote the paper.

This work was supported by the National Institute of Mental Health Grant R01 MH080716 to P.J.W. and Grant F31 MH090672 to M.J.K. We thank Daisy A. Burr for technical assistance.

The authors declare no competing financial interests.

Correspondence should be addressed to Dr. M. Justin Kim, Department of Psychology and Neuroscience, Duke University, 2020 West Main Street, Suite 0030, Durham, NC 27705. E-mail: justin.kim@duke.edu.

DOI:10.1523/JNEUROSCI.1375-17.2017

 0270-6474/17/379510-09$15.00/0

tion is useful in that it may be able to reconcile previous findings showing a purported valence effect because of a general tendency for negative stimuli to be rated as being higher in arousal compared with positive stimuli (Ito et al., 1998; Vaish et al., 2008).

Of critical importance, however, is the correlated nature of valence and arousal ratings. Specifically, emotional stimuli with clear valence (anger, happy, etc.) show a strong correlation between these dimensions, such that increased absolute valence intensity corresponds to increased arousal (Mattek et al., 2017). Consequently, studies contrasting positive/negative conditions with a neutral condition cannot delineate an effect of arousal from valence intensity. To distinguish an effect of valence intensity from an effect of arousal, there must be a non-neutral valence-ambiguous condition (i.e., where absolute valence intensity is "low" but arousal is still "high") (Mattek et al., 2017). Surprised facial expressions offer two unique advantages in this regard: (1) the effects of valence and arousal can be explored unconfounded by one another; and (2) negative and positive valence can be directly contrasted without resorting to the use of two or more distinct categories of emotional stimuli, eliminating potential confounds associated with cross-categorical emotions.

It is worth noting that, while previous studies have typically assumed that the valence of all surprised faces is equally ambiguous, photos of surprised faces that are available in standardized facial expression sets actually display varying degrees of valence ambiguity. That is, some individuals pose surprised facial expressions that are consistently rated as more "positive" (e.g., wonderment), other individuals pose surprised facial expressions that are perceived as more "negative" (e.g., shock), and still other individuals pose surprised facial expressions that are more ambiguous with respect to valence, consistently being rated within the center of the valence dimension. We hypothesize that this valence dimension can be decoded from visual features extracted from the facial features present within this single facial expression category. Thus, the purposes of the current study are the following: (1) to quantify any systematic agreement in the valence of surprised faces (i.e., surprised face stimuli that are consistently rated as negative or positive) contained within standardized databases that are widely used for scientific research; (2) to test whether the human amygdala tracks affective valence in particular, separate from arousal or emotion category; and (3) to link valence values to specific visual features extracted from the facial exemplars.

Materials and Methods

Participants

The present study was divided into four parts: (1) valence rating study, (2) emotion labeling study, (3) facial feature measurement, and (4) fMRI study. Independent groups of volunteers were recruited for the valence rating study, emotion labeling study, and fMRI study. For the valence rating study, 40 healthy Dartmouth College undergraduate students were recruited. Data from two participants were not recorded and thus removed from the analysis. As a result, a total of 38 participants (23 females; ages 18–21 years, mean age 18.7 years) were included in the final analysis. For the emotion labeling study, 17 healthy Dartmouth College undergraduate students volunteered (13 females; ages 18–20 years, mean age 18.8 years). For the fMRI study, 27 healthy Dartmouth College undergraduate students were recruited. Five individuals were excluded from all analyses due to excessive head movement during the scanning sessions (>1.5 mm). Thus, a total of 22 participants (12 females, ages 18–22 years, mean age 19.3 years) were included in the fMRI study. All participants were screened for past or current psychiatric illnesses (Axis I or II) using the Structured Clinical Interview for DSM-IV (First et al., 1995). No participants had any history of taking psychotropic medications. Before the experiment, all participants gave written, informed consent in accordance with the guidelines set by the Committee for the Protection of Human Subjects at Dartmouth College.

Valence rating study

Each participant saw a total of 63 surprised faces, which were selected from the Ekman (Ekman and Friesen, 1976), NimStim (Tottenham et al., 2009), and Karolinska facial expression sets (Lundqvist et al., 1998). Specifically, 12 faces from the Ekman stimulus set (6 females), 18 from the NimStim stimulus set (9 females), and 33 faces from the Karolinska stimulus set (19 females) were used. To keep the race/ethnicity consistent across the three different facial expression datasets (as the Ekman dataset only included white face stimuli), all of the stimuli included in the present study were limited to photos of white individuals. We note that the majority (68%) of the participants across all studies were also white. All faces were grayscaled and normalized for size and luminance. All faces were presented in a random order on a computer screen (visual angle $5° \times 8°$), using E-Prime software. Each face was presented once in a single run, which consisted of a total of 63 trials. On each trial, a single surprised face was presented on the computer screen for 1000 ms, followed by a two alternative forced-choice affective valence rating trial asking whether the face was positive or negative for 2000 ms. Behavioral ratings were collapsed across all 38 participants and analyzed for each of the 63 surprised face identities. The degree of ambiguity in affective valence for each stimulus face was operationally defined as the ratio of negative ratings each face received.

Emotion labeling study

While all 63 faces used in the current study were selected based on normative data that indicate each face was reliably categorized as surprise, considering the existence of consensus positive and consensus negative ratings (see Results section of the valence rating study for details), there was a need to confirm whether some of these faces were being consistently mistaken for any other emotion category than surprise. To this end, labeling accuracy data were collected from an additional 17 healthy volunteers (13 females; ages 18–20 years, mean age 18.8 years). Participants were presented with all 63 surprised faces in a random order on a computer screen (visual angle $5° \times 8°$); and for each trial, they were asked to select the emotion category that best describes the expression on the face. There were a total of 7 choices that the participants could select from (angry, disgust, fearful, happy, sad, surprised, neutral). Faces remained on the screen until the participants selected a response.

Facial feature measurements

Procedure. Consensus positive ($n = 12$) and negative ($n = 12$) surprised faces from the three facial expression sets were selected based on the results of the valence rating study. Specifically, for each valence type, there were 2 Ekman, 4 NimStim, and 6 Karolinska faces (matched for sex). For each identity, a total of four features were selected a priori for measurement, focusing on the eye and the mouth regions according to a rich body of literature suggesting that these regions provide critical affective information (Jack et al., 2014). These features were also consistent with participants' subjective report, when they were asked after the experiment to clarify which part of the face contributed to their valence decisions. Importantly, the majority of their responses were generally vague but showed a tendency to be focused on the features of the eyes and the mouth region (e.g., "something in the eyebrows," "maybe looked like smiling," etc.). The selected features were as follows: (1) distance between the upper eyelid and the eyebrows (Eye A); (2) distance between the upper and lower eyelids (Eye B); (3) distance between the labial commissures of the lips and the bottom of the chin (Mouth A); and (4) distance between the upper and lower lips (Mouth B). For features Eye A, Eye B, and Mouth A, an average of the measurements taken from the left and right side of the face was used. To account for potential differences in face size across identities, each individual measurement was normalized by dividing the vertical length of the whole face. All measurements were performed on a LCD computer monitor with a screen resolution of 1440×900 pixels, and the measuring units were in pixel counts. To rule out potential experimenter bias, two raters (M.J.K., J.S.) independently performed all measurements, and the interrater reliability was calculated using Cronbach's α, as well as the intraclass correlation coefficient. The

intraclass correlation analysis was performed using Model 2 (two-way random model), and measurements were assessed using absolute agreement across raters because the measurement error between raters here should not be systematic.

Machine learning classification analysis. For this procedure, we adopted a machine learning classification approach from the affective computing literature, which is in line with computational methods used to detect emotional content from facial expressions (Picard, 2010; Martinez and Valstar, 2015; e.g., D'Mello and Kory, 2015). All measurements were first scaled to be mean-centered and have unit variance, and then fed into a machine learning classifier, which used this information to distinguish positive versus negative surprised faces. To this end, the *scikit-learn* machine learning package (RRID:SCR_002577) in Python was used (Pedregosa et al., 2011). A support vector machine (SVM) with a linear kernel was selected for classification analysis. A linear SVM is a supervised learning algorithm that uses labeled training data to define a hyperplane that separates multidimensional space into two or more classes (Boser et al., 1992). Because there are two types of surprise being investigated here (12 positive and 12 negative surprise), classification accuracy at chance level was 50%. Hyperparameter optimization was achieved using a grid search algorithm; and through this analysis, a regularization parameter of $C = 0.1$ was used. A sixfold cross-validation was applied to the data such that the surprised faces are randomly and iteratively split into training and testing datasets. Specifically, a linear SVM classifier was trained iteratively on 5 folds and then tested on the remaining fold. Overall classification performance was then calculated by averaging the classification accuracy across folds. Feature weights were extracted for each of the four features to assess the relative importance of each feature to the trained classifier. In linear SVM, absolute values of the feature weights can be interpreted as how much each feature contributes to classification (Guyon and Elisseeff, 2003). Finally, to confirm the generalizability of the trained classifier, an additional 8 surprised faces (4 consensus positive, 4 consensus negative), which were never included in the cross-validation classifier training phase, were selected from the original 63 faces from the valence rating study as a validation set. The trained classifier performed classification on these 8 surprised faces, and its accuracy was measured.

Face averaging and subtraction. To illustrate the overall characteristics of the facial features associated with the consensus positive and negative surprised faces, following the machine learning classification analysis, the face stimuli were averaged separately for each valence type using an online software (http://www.facefacts.scot; Institute of Neuroscience and Psychology, University of Glasgow). The landmarks used to average the faces were delineated semimanually; first, the software requires a manual input for three major landmarks (left eye, right eye, mouth). Based on these inputs, the software automatically places markers for various specific landmarks (e.g., eyebrows, outer upper lip) in their approximate locations. Then we manually tweaked these markers to best match the actual facial structure. The landmarks of each individual face were carefully inspected to maximize the quality of the averaging algorithm. Using an approach similar to the methods described by Ahs et al. (2014), the averaged faces were subsequently transformed to *z* scores and then subtracted from one another using Python.

fMRI study

Experimental design. Consensus positive ($n = 12$) and negative ($n = 12$) surprised faces were taken from the facial feature measurement study. An additional "ambiguous" sample of surprised faces ($n = 12$) that displayed a ratio of negative ratings between the consensus positive and negative face stimuli was selected balancing for the different facial expression datasets and sex. Using E-Prime software, all stimuli were back projected onto a screen (visual angle 5° × 8°), on which the participants viewed during fMRI scanning using a mirror that was mounted on the head coil. All faces were grayscaled and normalized for size and luminance. A passive viewing paradigm that has been shown to reliably elicit amygdala activity to facial expressions of emotion was adapted for the current study (Kim and Whalen, 2009; for meta-analysis and review, see Costafreda et al., 2008). During fMRI scanning sessions, participants viewed two runs consisting of 18 s blocks of consensus positive, consensus negative, and ambiguous surprised faces, interleaved with 18 s blocks showing a single fixation crosshair at the center of the screen (e.g., example run structure: +P+A+N+A+N+P+N+P+A+). Within each block, surprised faces were presented for 200 ms followed by a 300 ms interstimulus interval, yielding a total of 36 trials per block. The order of the faces within each block was pseudorandomized to ensure that the same face was not presented more than twice in a row. There were three blocks for each condition in each run, and participants were scanned for the duration of two runs, which lasted for a total of 12 min 20 s.

Image acquisition. All participants were scanned at the Dartmouth Brain Imaging Center using a 3.0 Tesla Philips Intera Achieva Scanner (Philips Medical Systems) equipped with an 8-channel head coil. High-resolution anatomical T1-weighted images were collected using a high-resolution 3D MP-RAGE, with 160 contiguous 1-mm-thick sagittal slices (TE = 4.6 ms, TR = 9.8 ms, FOV = 240 mm, flip angle = 8°, voxel size = 1 × 0.94 × 0.94 mm). Functional images were acquired using echo-planar T2*-weighted imaging sequence. Each volume consisted of 36 interleaved 3-mm-thick slices with 0.5 mm interslice gap (TE = 35 ms, TR = 2000 ms, FOV = 240 mm, flip angle = 90°, voxel size = 3 × 3 × 3.5 mm).

fMRI data analysis. All fMRI data were preprocessed using SPM12 (RRID:SCR_007037). Raw functional images were corrected for head movement. None of the remaining 22 participants had head movement >1.5 mm in any direction. Functional images were then normalized to standard space (3 × 3 × 3 mm) using the MNI-152 template. Spatial smoothing was applied to the normalized functional images using a Gaussian kernel of 6 mm FWHM. By using a boxcar function convolved with a HRF and covariates of no interests (a session mean, a linear trend for each run to account for low-frequency drift and six movement parameters derived from realignment corrections), contrast maps for negative versus positive surprise (our a priori planned contrast) as well as each condition versus fixation were generated for each participant. Contrast maps were then entered into a random effects model, which accounts for intersubject variability and allows population based inferences to be drawn.

Statistical analyses. Given our a priori hypothesis of the amygdala, we imposed a significance threshold of $p < 0.05$ corrected for multiple comparisons over the bilateral amygdala volume defined using the Automated Anatomical Labeling atlas (Maldjian et al., 2003), as determined by Monte Carlo simulations ($n = 10{,}000$) implemented in *3dClustSim*, in conjunction with *3dFWHMx* and an updated method (autocorrelation function) to estimate smoothness in the data, within AFNI software (RRID:SCR_005927) (Cox, 1996). The corrected $p < 0.05$ corresponded to uncorrected $p < 0.005$ and $k \geq 4$ voxels (108 mm^3). We note that the results remained consistent (cluster size of the significant left amygdala voxels increased from 14 to 22 voxels and right amygdala voxels increased from 2 to 15 voxels; see Results) when nonparametric permutation tests ($n = 10{,}000$) were performed on the data to determine significant voxels at $p < 0.05$ corrected for multiple comparisons, using *randomize* along with the threshold-free cluster enhancement method implemented in FSL (RRID:SCR_002823) (Smith and Nichols, 2009; Winkler et al., 2014). For all other brain regions, whole-brain corrected $p < 0.05$ threshold was achieved by using uncorrected $p < 0.001$, $k \geq 77$ voxels (2079 mm^3) through the use of Monte Carlo simulations described above.

Postscan assessment. Upon exiting the scanner, participants were presented with the same blocks of surprised faces that they have seen in the scanner in a random order on a computer screen. Similar to the methods described by Kim et al. (2003), after each block, they were instructed to rate the overall valence and arousal of each block on a 9-point Likert scale. Valence ratings were collected using a scale that ranged from 1 to 9 (positive to negative), which was subsequently converted to a scale that ranged from −4 to 4 (negative to positive; the scale itself was an identical 9-point Likert scale). Each individual's average valence and arousal ratings for the negative, ambiguous, and positive surprise blocks were then calculated. Average valence ratings for each type of surprise were used as a manipulation check: to ensure that the participants in the fMRI study have indeed viewed the negative/ambiguous/positive surprised faces in a similar manner as the participants in the valence rating study.

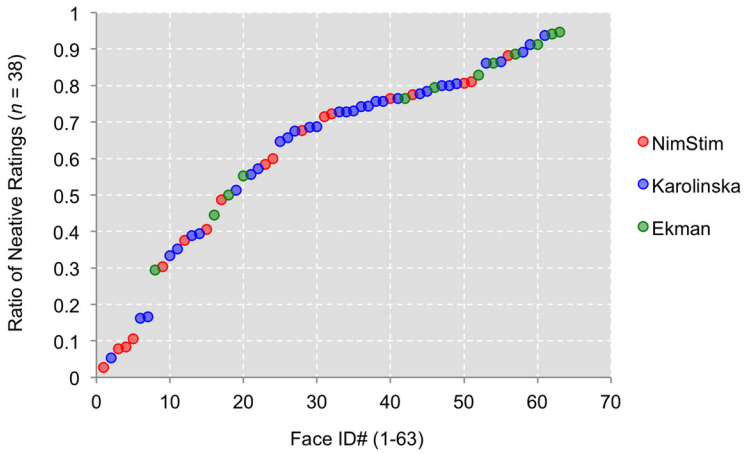

Figure 1. Scatterplot depicting the distribution of emotional ambiguity in 63 surprised faces, taken from the Ekman (green), the NimStim (red), and the Karolinska (blue) facial expressions sets, sorted by the ratio of negative ratings.

As a final step, participants filled out the State-Trait Anxiety Inventory Form-Y (STAI) (Spielberger et al., 1988) and the Positive and Negative Affective Schedule (PANAS) (Watson et al., 1988). All postscan measurements were used to quantify concurrent mood and anxiety levels that may influence the perceived valence of surprised faces. State anxiety scores (STAI-S) were used as a proxy for each participant's anxiety levels. Both PANAS negative and positive mood scores were used as an index of concurrent mood.

Results

Valence rating data

Calculating the consensus ratio of negative ratings for each surprised face revealed a varying degree of ambiguity across all 63 faces, ranging from as low as 0.03 (consensus positive) to as high as 0.95 (consensus negative; see Fig. 1; Table 1). Current results provide evidence that the valence of surprised faces, when perceived devoid of contextual cues, can display a wider range of valence ambiguity than previously assumed. From these 63 surprised faces, 12 consensus positive and 12 consensus negative identities, matched for sex (6 males, 6 females) and stimulus set (2 Ekman, 4 Nimstim, 6 Karolinska) per valence, were selected and used for subsequent studies.

Emotion labeling accuracy data

Results showed that surprised faces with a higher ratio of negative ratings were more likely to be confused with fearful faces than any other emotional expressions (Fig. 2). However, those consensus negative surprised faces were still categorized as surprised overall (i.e., in the "worst" case, the face was categorized as 59% surprised, 29% fearful, 6% happy, and 6% sad). Finally, one consensus positive face tended to be confused with happy but was still primarily categorized as surprised (59% surprised, 35% happy, 6% fearful).

Facial feature measurements

Inter-rater reliability

For all four features, interrater reliability was high, as indicated by Cronbach's α (Eye A = 0.97; Eye B = 0.88; Mouth A = 0.89; Mouth B = 0.98). These results were further corroborated in that all four facial features achieved high intraclass correlation coefficients (Eye A = 0.83; Eye B = 0.79; Mouth A = 0.79; Mouth B = 0.95; all p values <0.00001). Thus, for the subsequent classification analysis, measurements from one rater (M.J.K.) were used.

Classification accuracy data

Cross-validation tests showed that the linear SVM classifier was able to correctly classify negative and positive surprised faces at 75% accuracy (SEM 12.9%) when trained with the four features. Feature weights indicated that this classifier relied on information from the mouth regions (Mouth A = 0.32; Mouth B = 0.55) relative to the eye regions (Eye A = −0.28; Eye B = −0.12). The absolute value of the feature weights in a linear SVM classification analysis corresponds to the relative importance of each feature, yielding a rank-order of Mouth B > Mouth A > Eye A > Eye B. A classification accuracy of 87.5% was achieved when a validation set of 8 novel surprised faces were used (1 misclassification each of a total of 8 faces). Figure 3 summarizes the outcomes from the classifier analysis.

Averaged faces

Averaged (i.e., composite) positive and negative surprised faces are presented in Figure 4. Upon visual inspection, the averaged positive surprised face most notably differed in the mouth region displaying a more prominent jaw drop. We note that, while the areas around the hair and below the ears include substantial variability across individual faces, our four measurements were unaffected by the noise in those regions. For future studies that may include facial features around the noisy areas, further manipulation of the stimuli (e.g., removing the hair) may be necessary.

fMRI data

Neuroimaging data

The consensus negative versus positive surprise contrast revealed significantly increased activity in the left amygdala (MNI −24, −6, −18, $t_{(21)} = 4.61$, $z = 3.79$, $k = 14$ voxels; Cohen's $d = 1.07$; corresponding to $p < 0.05$ corrected for multiple comparisons within the amygdala; Fig. 5). Power analysis was performed to compute the power for the amygdala, using a standard effect size that is associated with typical emotional tasks (Cohen's d = ~0.6) (Poldrack et al., 2017), α level, and sample size. Given these parameters, the power to detect significant differences in the amygdala was 77%. Amygdala activity was not correlated with either concurrent anxiety (STAI-S scores) or mood (PANAS-P, PANAS-N scores; all p values >0.05). *Post hoc* analyses showed that left amygdala activity to ambiguous surprise was not significantly different from either negative or positive surprise (both p values >0.05). No clusters within the amygdala displayed increased activity as a function of arousal when the negative and positive > ambiguous surprise contrast was examined *post hoc*, according to the postscan behavioral data. These findings highlight the fact that the amygdala is responsive to a negative valence signal embedded within surprised faces, likely comprising a configuration of certain facial features. No other brain regions showed significant differential activity across negative, ambiguous, and positive surprise face conditions.

Postscan behavioral data

Repeated-measures ANOVA demonstrated that, as expected, participants rated the affective valence of the surprised face blocks in accordance with their predetermined categories ($F_{(2,42)} = 64.79$, $p < 0.000001$). *Post hoc* analysis showed that

Table 1. Ratio of negative ratings for all 63 surprised faces

Face ID	Ratio of negative ratings	Face ID	Ratio of negative ratings
Tottenham et al. (2009)			
01F	0.68	20M	0.08
02F	0.30	23M	0.81
03F	0.11	27M	0.38
05F	0.03	28M	0.71
06F	0.76	32M	0.72
07F	0.58	34M	0.41
08F	0.81	35M	0.88
09F	0.77	36M	0.60
10F	0.08	37M	0.49
Lundqvist et al. (1998)			
AF01	0.35	AF32	0.39
AF02	0.74	AF34	0.73
AF03	0.76	AM02	0.73
AF09	0.56	AM03	0.80
AF11	0.86	AM05	0.74
AF12	0.17	AM06	0.69
AF13	0.05	AM09	0.86
AF16	0.76	AM11	0.39
AF19	0.68	AM12	0.78
AF20	0.80	AM13	0.33
AF21	0.78	AM18	0.81
AF22	0.16	AM20	0.94
AF23	0.51	AM24	0.66
AF24	0.76	AM31	0.57
AF26	0.89	AM34	0.65
AF27	0.91	AM35	0.73
AF30	0.69		
Ekman and Friesen (1976)			
C	0.44	EM	0.79
MF	0.50	GS	0.55
NR	0.94	JJ	0.89
SW	0.83	PE	0.91
PF	0.29	WF	0.86
JM	0.76	JB	0.95

negative blocks (-1.36 ± 0.87) were significantly rated as more negative than ambiguous (-0.88 ± 0.74; $t_{(21)} = -2.11$, $p = 0.047$) or positive blocks (1.53 ± 1.03; $t_{(21)} = -9.28$, $p < 0.000001$). Participants also rated the positive blocks as being significantly more positive than the ambiguous blocks ($t_{(21)} = 8.88$, $p < 0.000001$; Fig. 6*A*).

Participants rated the arousal of the negative (4.42 ± 1.49) and positive (4.6 ± 1.39) surprised face blocks as significantly different from the ambiguous blocks (3.92 ± 1.44; $F_{(2,42)} = 5.31$, $p = 0.009$). *Post hoc* analysis confirmed that both negative ($t_{(21)} = 2.43$, $p = 0.024$) and positive ($t_{(21)} = 3.17$, $p = 0.005$) blocks received significantly higher arousal ratings than the ambiguous blocks. Importantly, no significant difference in arousal ratings between the negative and positive blocks was observed ($t_{(21)} = 0.77$, $p = 0.45$; Fig. 6*B*). No significant correlations were observed between the absolute valence and arousal ratings for the negative, ambiguous, or positive surprised faces.

Descriptive statistics of the self-report measures of anxiety and mood are as follows: STAI-S (33.18 ± 8.64), PANAS-P (37.95 ± 5), and PANAS-N (17.36 ± 2.77).

Discussion

Here, we tested whether the human amygdala is sensitive to differences in the valence value of surprised facial expressions. Given that valence and arousal ratings were uncorrelated (see also Mattek et al., 2017), we demonstrated a dimensional valence affect within the amygdala that was unconfounded by arousal ratings or expression category. In addition, we showed that a machine learning algorithm could reliably differentiate between positive and negative surprised expressions. These data allowed us to assert that the degree to which the mouth is open predicts a qualitatively more positive surprised expression. It follows then that the amygdala may be sensitive to this low-level visual signal that is associated with subjective valence for surprised facial expressions.

Human amygdala can encode valence unconfounded by arousal

The amygdala responses characterized here are consistent with previous reports showing a relationship between stimulus valence and amygdala activity. A number of fMRI studies have found that amygdala BOLD response is greater to negative compared with neutral items (Williams et al., 2004; Chang et al., 2015), and still others have found that amygdala BOLD response is greater to negative compared with positive items (Anders et al., 2008; Lindquist et al., 2016). In comparison, some experiments find that the amygdala responds equally to positive and negative stimuli (Garavan et al., 2001), and still others find that the amygdala responds more to positive compared with negative stimuli (Sergerie et al., 2008). Importantly, these studies are unable to address whether categorical differences in the stimuli used to elicit amygdala response may be driving these effects.

Compared with other categories of facial expressions, the valence signal of surprised faces is subtle (i.e., limited range of responses to the positive/negative surprise on a 9-point Likert scale), yet reliable (i.e., existence of consensus positive/negative surprise). Considering these characteristics of surprise, the current fMRI finding speaks to the sensitivity of the human amygdala to valence information. An important detail here was that arousal was equivalent between positive and negative surprised faces. This is consistent with a recent fMRI study using a multivariate pattern analysis that reported the activity of the amygdala closely represented the valence, not arousal, of various odor stimuli (Jin et al., 2015).

The present fMRI data show that amygdala activity can capture subjective valence exclusively when experimental conditions remove subjective arousal as a potential confound. We are not suggesting that the amygdala does not also encode arousal; we of course assume the amygdala processes both dimensions and that the subnuclei of the amygdala may differ in their relative contributions to processing these two dimensions (for review, see Whalen et al., 2009). That being said, given reports showing that the amygdala can function to provide information about the salience or arousal value of environmental events (e.g., Wilson-Mendenhall et al., 2013), we thought it critical to offer the example of the surprised expression category to demonstrate that the amygdala can also track the valence of environmental events (Whalen, 1998; Belova et al., 2007).

An interesting parallel to the current fMRI study using surprised faces is a series of reports on the relationship between the perceived trustworthiness traits of emotionally neutral faces and amygdala activity. Similar to the consensus valence ratings of surprised faces, Engell et al. (2007) showed that amygdala activity tracked consensus ratings of trustworthiness of neutral faces, such that more untrustworthy faces elicited higher amygdala response. Furthermore, variability in trustworthiness ratings was identified as being closely related to a general valence dimension, and the amygdala activity increased as a function of this dimension

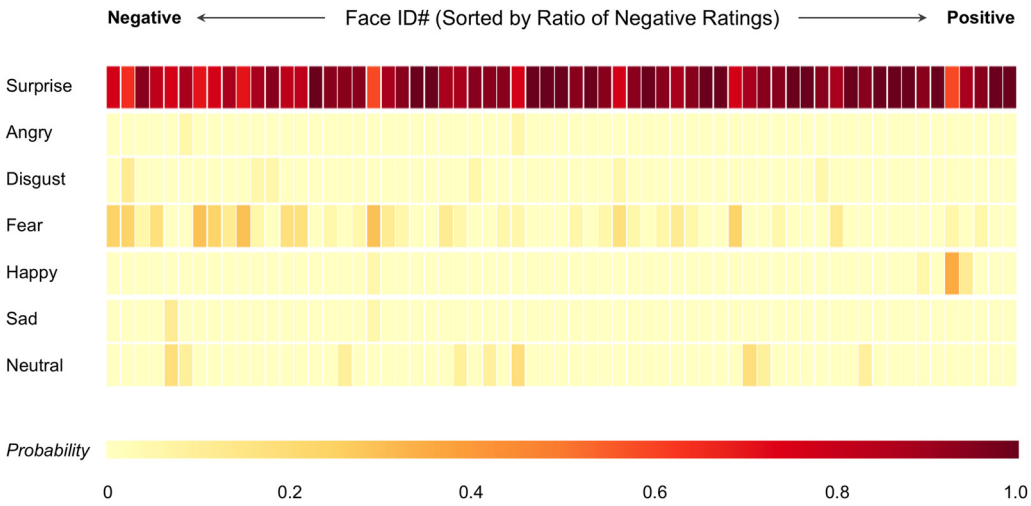

Figure 2. Heat map depicting the confusion matrix of all 63 surprised faces, sorted by the ratio of negative ratings. Color bar represents the probability of being categorized as the corresponding emotion. While all faces were primarily categorized as surprised, they were most frequently confused with fear.

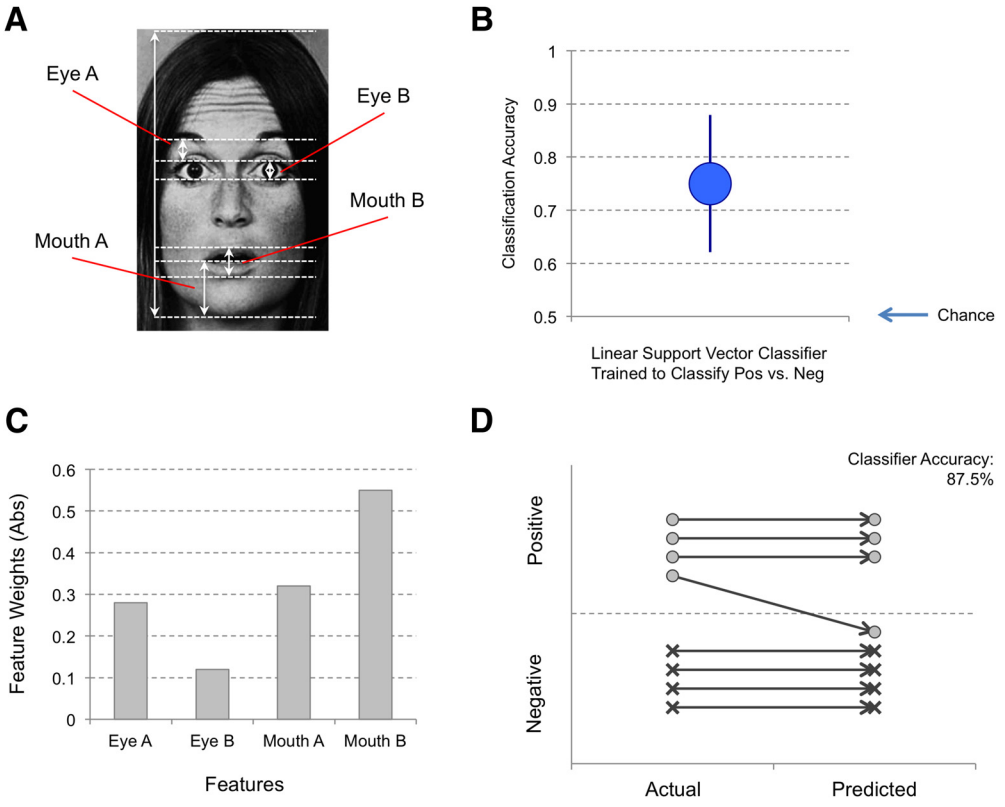

Figure 3. Summary of the classification results. ***A***, Four facial features that were measured for classification analysis. ***B***, A linear SVM classifier that was trained to distinguish negative versus positive surprised faces demonstrated 75% accuracy from cross-validation testing. Error bars indicate SEM. ***C***, Absolute values of the feature weights indicate that the mouth regions have more informational value than the eye regions for this trained classifier. ***D***, An additional validation test was performed on a holdout set of surprised faces that were never used for the cross-validation training of the classifier, and showed a classification accuracy of 87.5%.

(Todorov and Engell, 2008). In other words, higher amygdala activity was associated with more negative signals embedded within the facial features of otherwise neutral expressions, similar to the observations from the current study. Collectively, these converging findings suggest that the human amygdala is sensitive to valence information gleaned from the faces of others, based on certain features that may have likely served as biologically relevant predictive cues.

Feature-based valence signal within surprised faces guides amygdala responsivity

The current study also showed that a machine learning classifier could be trained to reliably discriminate positive and negative surprised faces, using a combination of four simple facial features. Upon examining the relative contributions of each of the four features on classifier performance, the two mouth regions were observed to contain more information about valence than

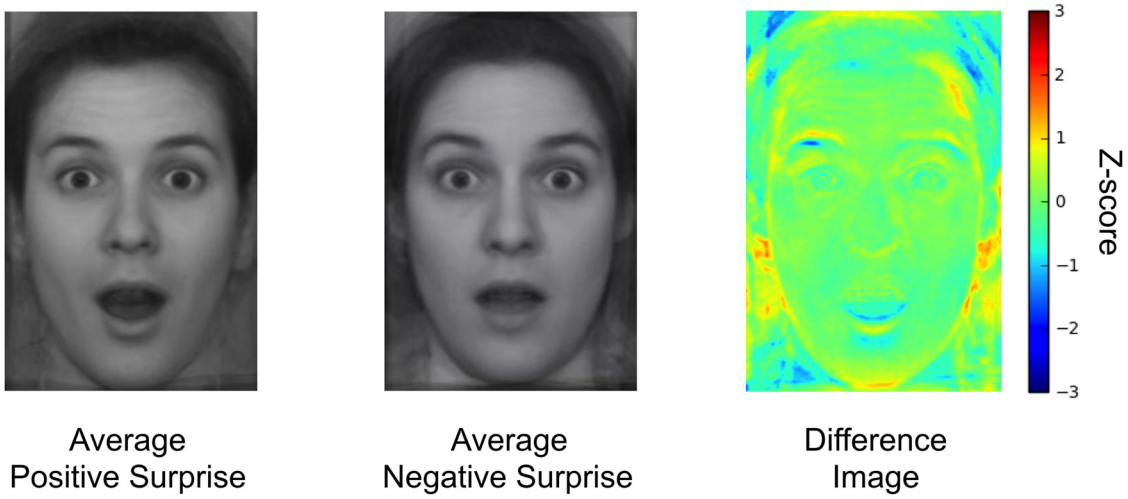

Figure 4. Average faces of consensus positive and consensus negative surprise, and a heat map depicting the differences between the two images. Largest differences were found in the mouth and the right eyebrow area.

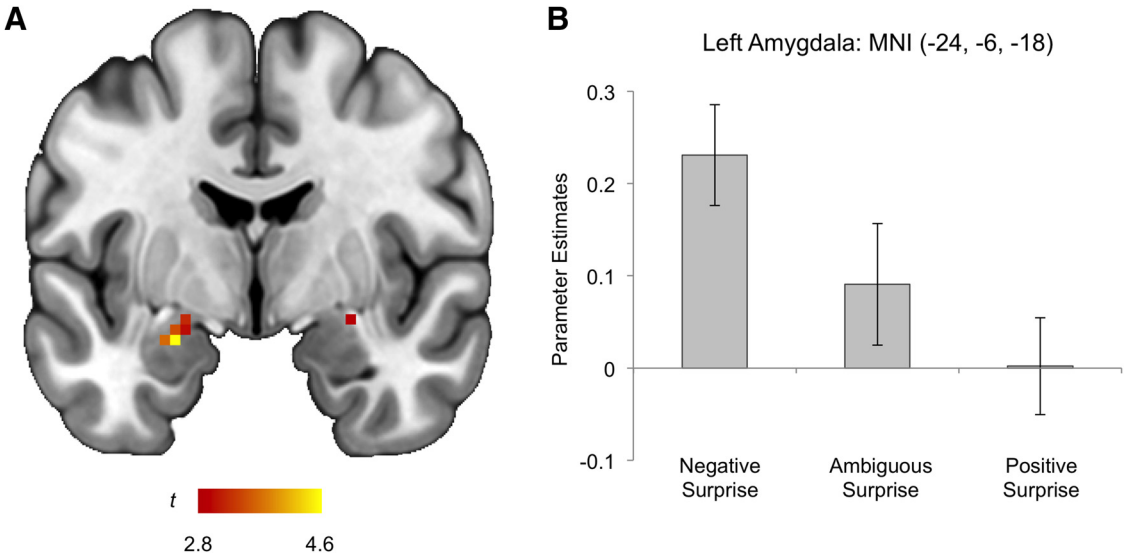

Figure 5. Summary of the fMRI results. ***A***, A coronal slice of the brain depicting the voxel clusters in the amygdala whose activity tracked the valence (negative > ambiguous > positive) of surprised faces. ***B***, Left amygdala activity showing a linear increase as a function of valence. Error bars indicate SEM.

the two eye regions. Specifically, a more open mouth in the vertical direction predicted more positive ratings. Given the high concordance rate for these consensus positive and negative surprised faces, it is reasonable to hypothesize that human observers may also extract information from these facial features in a similar manner. Interestingly, participants could not confidently pinpoint certain facial features that guided their affective valence decisions in the current experiment. Participants' generally vague responses (e.g., "just feels negative," "something in the eyebrows"), combined with the fact that there still was a high degree of agreement on the consensus positive and negative surprised faces, are consistent with the view that these facial features are being processed implicitly (Farah et al., 1998). Interestingly, the mouth region was recently suggested to be more informative than the eye region in discriminating facial expressions across multiple categories of emotion (Blais et al., 2012). Our data expand this report by showing the utility of the mouth region in distinguishing positive versus negative faces within a single emotional category.

The present results converge with a previous study of composite facial expression categories. Du et al. (2014) showed that an algorithm could distinguish between composite facial expressions when human actors were asked to deliberately combine two basic emotional expressions, such as surprised faces blended with either happy or fearful faces. The present results critically extend this work to show that there is sufficient variability of facial muscle expression, within the category of surprised expressions available across standardized facial expression stimulus sets, for an algorithm to discriminate this valence dimension. Another study used computer-generated dynamic facial expressions (i.e., a short video clip of a neutral face morphing into either a positive or negative surprised face) that the participants viewed while their brain activity was recorded in an fMRI scanning session, and found increased amygdala activity to positive versus negative surprise (Vrticka et al., 2014). There are several key differences between studies, which include the use of naturalistic versus artificial face stimuli and static photos versus movies of facial expressions. More importantly, unlike Vrticka et al. (2014), the

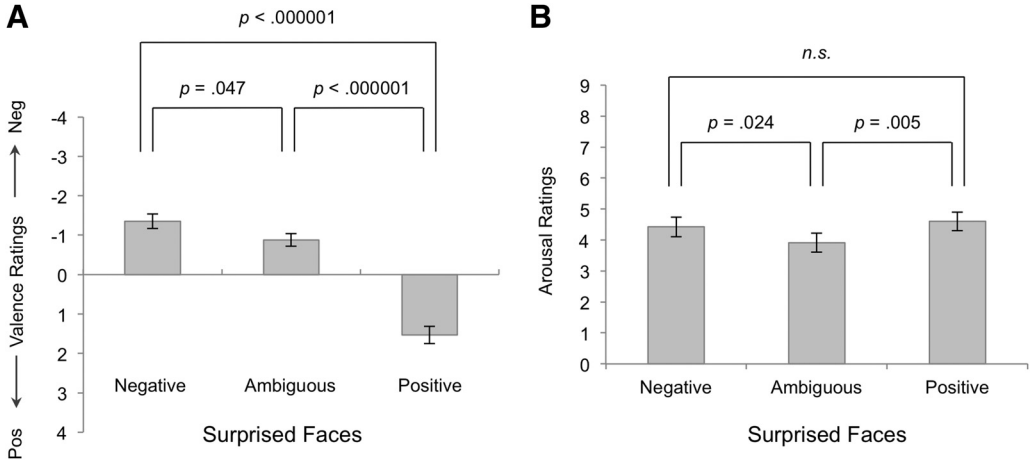

Figure 6. Postscan behavioral ratings of valence and arousal. ***A***, Significant difference in valence ratings between conditions was observed. ***B***, Arousal ratings were significantly lower for ambiguous surprise compared with negative or positive surprise. There were no significant differences in arousal ratings between negative and positive surprised faces. Error bars indicate SEM.

present study controlled for subjective levels of arousal between positive and negative surprise, and established that these positive and negative surprise faces were indeed perceived as the category of surprise overall, which may have contributed to the discrepancy in the results across studies.

Finally, the findings reported here have implications for future research that would use these surprised expression faces as experimental stimuli. Depending the experimental goals, it would be important to stratify selection of specific stimuli based on the designation of positive, negative, or ambiguous surprise. That is, the degree of emotional ambiguity should be quantified and then balanced (e.g., equal numbers of positive, negative, and ambiguous exemplars) or unbalanced (e.g., use only the ambiguous exemplars) depending on the study design. Additionally, the present data could inform future research that the amygdala's responsiveness to specific facial features may depend on context. For example, studies have shown that the amygdala is sensitive to the eye region when comparing fearful versus happy facial expressions in a backward masking paradigm (Straube et al., 2010; Kim et al., 2016), which is seemingly in contrast with the present data. We suggest that context is key: the amygdala being more sensitive to the mouth region within the context of surprise, and to the eye region within the context of fear (as well as backward masking), could be understood through a general framework that suggests that the amygdala is most sensitive to facial features that offer useful, biologically relevant information in a given context. As such, we speculate that the human amygdala is sensitive to the mouth within the category of surprise because it is the best available source of biologically relevant information (i.e., valence signal) when comparing negative versus positive surprised faces. It remains an open question as to whether the amygdala is sensitive to the mouth in other expressions. Future studies might seek to systematically manipulate the mouth region, use an algorithm that can automatically detect relevant facial features, or delineate low versus high spatial frequency components of surprised faces (e.g., Méndez-Bértolo et al., 2016) to address this issue.

In conclusion, the current study demonstrated that human amygdala responses track affective valence, when holding emotional arousal and emotion category constant. This was achieved by focusing on the emotion category of surprise, single facial expression category that can capture the dimension of valence. Differences in the configuration of facial features predicted these differences in valence, with more vertical opening of the mouth being most predictive of a qualitatively more positive surprised expression. Together, these findings suggest that the human amygdala is responsive to very subtle valence-related cues embedded within facial features.

References

Adolphs R (2010) What does the amygdala contribute to social cognition? Ann N Y Acad Sci 1191:42–61. CrossRef Medline

Ahs F, Davis CF, Gorka AX, Hariri AR (2014) Feature-based representations of emotional facial expressions in the human amygdala. Soc Cogn Affect Neurosci 9:1372–1378. CrossRef Medline

Anders S, Eippert F, Weiskopf N, Veit R (2008) The human amygdala is sensitive to the valence of pictures and sounds irrespective of arousal: an fMRI study. Soc Cogn Affect Neurosci 3:233–243. CrossRef Medline

Anderson AK, Phelps EA (2001) Lesions of the human amygdala impair enhanced perception of emotionally salient events. Nature 411:305–309. CrossRef Medline

Anderson AK, Christoff K, Stappen I, Panitz D, Ghahremani DG, Glover G, Gabrieli JD, Sobel N (2003) Dissociated neural representations of intensity and valence in human olfaction. Nat Neurosci 6:196–202. CrossRef Medline

Belova MA, Paton JJ, Morrison SE, Salzman CD (2007) Expectation modulates neural responses to pleasant and aversive stimuli in primate amygdala. Neuron 55:970–984. CrossRef Medline

Blais C, Roy C, Fiset D, Arguin M, Gosselin F (2012) The eyes are not the window to basic emotions. Neuropsychologia 50:2830–2838. CrossRef Medline

Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. In: 5th Annual Association for Computing Machinery workshop on computational learning theory (Haussler D, ed), pp 144–152. Pittsburgh: ACM.

Breiter HC, Etcoff NL, Whalen PJ, Kennedy WA, Rauch SL, Buckner RL, Strauss MM, Hyman SE, Rosen BR (1996) Response and habituation of the human amygdala during visual processing of facial expression. Neuron 17:875–887. CrossRef Medline

Chang LJ, Gianaros PJ, Manuck SB, Krishnan A, Wager TD (2015) A sensitive and specific neural signature for picture-induced negative affect. PLoS Biol 13:e1002180. CrossRef Medline

Costafreda SG, Brammer MJ, David AS, Fu CH (2008) Predictors of amygdala activation during the processing of emotional stimuli: a meta-analysis of 385 PET and fMRI studies. Brain Res Rev 58:57–70. CrossRef Medline

Cox RW (1996) AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput Biomed Res 29:162–173. CrossRef Medline

D'Mello S, Kory J (2015) A review and meta-analysis of multimodal affect detection systems. ACM Comput Surv 47:Article 43.

Du S, Tao Y, Martinez AM (2014) Compound facial expressions of emotion. Proc Natl Acad Sci U S A 15:E1454–E1462. CrossRef Medline

Ekman PF, Friesen WV (1976) Pictures of facial affect. Palo Alto, CA: Consulting Psychologists.
Engell AD, Haxby JV, Todorov A (2007) Implicit trustworthiness decisions: automatic coding of face properties in the human amygdala. J Cogn Neurosci 19:1508–1519. CrossRef Medline
Farah MJ, Wilson KD, Drain M, Tanaka JN (1998) What is "special" about face perception? Psychol Rev 105:482–498. CrossRef Medline
First M, Spitzer M, Williams J, Gibbon M (1995) Structured clinical interview for DSM-IV (SCID). Washington, DC: American Psychiatric Association.
Fitzgerald DA, Angstadt M, Jelsone LM, Nathan PJ, Phan KL (2006) Beyond threat: amygdala reactivity across multiple expressions of facial affect. Neuroimage 30:1441–1448. CrossRef Medline
Garavan H, Pendergrass JC, Ross TJ, Stein EA, Risinger RC (2001) Amygdala response to both positively and negatively valenced stimuli. Neuroreport 12:2779–2783. CrossRef Medline
Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182.
Ito TA, Cacioppo JT, Lang PJ (1998) Eliciting affect using the International Affective Picture System: trajectories through evaluative space. Pers Soc Psychol Bull 24:855–879. CrossRef
Jack RE, Garrod OG, Schyns PG (2014) Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. Curr Biol 24:187–192. CrossRef Medline
Jin J, Zelano C, Gottfried JA, Mohanty A (2015) Human amygdala represents the complete spectrum of subjective valence. J Neurosci 35:15145–15156. CrossRef Medline
Kim H, Somerville LH, Johnstone T, Alexander AL, Whalen PJ (2003) Inverse amygdala and medial prefrontal cortex responses to surprised faces. Neuroreport 14:2317–2322. CrossRef Medline
Kim MJ, Whalen PJ (2009) The structural integrity of an amygdala-prefrontal pathway predicts trait anxiety. J Neurosci 29:11614–11618. CrossRef Medline
Kim MJ, Solomon KM, Neta M, Davis FC, Oler JA, Mazzulla EC, Whalen PJ (2016) A face versus non-face context influences amygdala responses to masked fearful eye whites. Soc Cogn Affect Neurosci 11:1933–1941. CrossRef Medline
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. CrossRef Medline
Lundqvist D, Flykt A, Ohman A (1998) The Karolinska directed emotional faces-KDEF [CD-ROM]. Stockholm: Department of Clinical Neuroscience, Psychology section, Karolinska Institutet.
Maldjian JA, Laurienti PJ, Kraft RA, Burdette JH (2003) An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI datasets. Neuroimage 19:1233–1239. CrossRef Medline
Martinez B, Valstar MF (2015) Advances, challenges, and opportunities in automatic facial expression recognition. In: Advances in face detection and facial image analysis (Kawulok M, Celebi E, Smolka B, eds), pp 63–100. New York: Springer.
Mattek AM, Wolford GL, Whalen PJ (2017) A mathematical model captures the structure of subjective affect. Perspect Psychol Sci 12:508–526. CrossRef Medline
Méndez-Bértolo C, Moratti S, Toledano R, Lopez-Sosa F, Martínez-Alvarez R, Mah YH, Vuilleumier P, Gil-Nagel A, Strange BA (2016) A fast pathway for fear in human amygdala. Nat Neurosci 19:1041–1049. CrossRef Medline
Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830.
Picard R (2010) Affective computing: from laughter to IEEE. IEEE Affect Comput 1:11–17. CrossRef
Poldrack RA, Baker CI, Durnez J, Gorgolewski KJ, Matthews PM, Munafò MR, Nichols TE, Poline JB, Vul E, Yarkoni T (2017) Scanning the horizon: towards transparent and reproducible neuroimaging research. Nat Rev Neurosci 18:115–126. CrossRef Medline
Russell JA (1980) A circumplex model of affect. J Pers Soc Psychol 39:1161–1178. CrossRef
Sergerie K, Chochol C, Armony JL (2008) The role of the amygdala in emotional processing: a quantitative meta-analysis of functional neuroimaging studies. Neurosci Biobehav Rev 32:811–830. CrossRef Medline
Smith SM, Nichols TE (2009) Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. Neuroimage 44:83–98. CrossRef Medline
Spielberger CD, Gorsuch RL, Lushene RE (1988) STAI-Manual for the state trait anxiety inventory. Palo Alto, CA: Consulting Psychologists.
Straube T, Dietrich C, Mothes-Lasch M, Mentzel HJ, Miltner WH (2010) The volatility of the amygdala response to masked fearful eyes. Hum Brain Mapp 31:1601–1608. CrossRef Medline
Todorov A, Engell AD (2008) The role of the amygdala in implicit evaluation of emotionally neutral faces. Soc Cogn Affect Neurosci 3:302–312. CrossRef Medline
Tomkins SS, McCarter R (1964) What and where are the primary affects? Some evidence for a theory. Percept Mot Skills 18:119–158. CrossRef Medline
Tottenham N, Tanaka JW, Leon AC, McCarry T, Nurse M, Hare TA, Marcus DJ, Westerlund A, Casey BJ, Nelson C (2009) The NimStim set of facial expressions: judgments from untrained research participants. Psychiatry Res 168:242–249. CrossRef Medline
Vaish A, Grossmann T, Woodward A (2008) Not all emotions are created equal: the negativity bias in social-emotional development. Psychol Bull 134:383–403. CrossRef Medline
Vrticka P, Lordier L, Bediou B, Sander D (2014) Human amygdala response to dynamic facial expressions of positive and negative surprise. Emotion 14:161–169. CrossRef Medline
Watson D, Clark LA, Tellegen A (1988) Development and validation of brief measures of positive and negative affect: the PANAS scales. J Pers Soc Psychol 54:1063–1070. CrossRef Medline
Whalen PJ (1998) Fear, vigilance, and ambiguity: initial neuroimaging studies of the human amygdala. Curr Dir Psychol Sci 7:177–188. CrossRef
Whalen PJ, Davis FC, Oler JA, Kim H, Kim MJ, Neta M (2009) Human amygdala responses to facial expressions of emotion. In: The human amygdala (Whalen PJ, Phelps EA, eds), pp 265–288. New York: Guilford.
Williams MA, Morris AP, McGlone F, Abbott DF, Mattingley JB (2004) Amygdala responses to fearful and happy facial expressions under conditions of binocular suppression. J Neurosci 24:2898–2904. CrossRef Medline
Wilson-Mendenhall CD, Barrett LF, Barsalou LW (2013) Neural evidence that human emotions share core affective properties. Psychol Sci 24:947–956. CrossRef Medline
Winkler AM, Ridgway GR, Webster MA, Smith SM, Nichols TE (2014) Permutation inference for the general linear model. Neuroimage 92:381–397. CrossRef Medline