OXFORD

Social Cognitive and Affective Neuroscience, 2022, **17**, 549–558

DOI: https://doi.org/10.1093/scan/nsab134
Advance Access Publication Date:23 December 2021
Original Manuscript

Localized task-invariant emotional valence encoding revealed by intracranial recordings

Daniel S. Weisholtz,[1] Gabriel Kreiman,[2] David A. Silbersweig,[1] Emily Stern,[1,4] Brannon Cha,[1,5] and Tracy Butler[3]

[1]Department of Neurology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA
[2]Boston Children's Hospital, Harvard Medical School, Boston, MA 02115, USA
[3]Department of Radiology, Weill Cornell Medical Center, New York 10065, USA
[4]Present address: Ceretype Neuromedicine, Inc.
[5]Present address: University of California San Diego School of Medicine.
Correspondence should be addressed to Daniel S. Weisholtz, Department of Neurology, Brigham and Women's Hospital, Harvard Medical School, 60 Fenwood Road, Boston, MA 02115, USA. E-mail: dweisholtz@partners.org.

Abstract

The ability to distinguish between negative, positive and neutral valence is a key part of emotion perception. Emotional valence has conceptual meaning that supersedes any particular type of stimulus, although it is typically captured experimentally in association with particular tasks. We sought to identify neural encoding for task-invariant emotional valence. We evaluated whether high-gamma responses (HGRs) to visually displayed words conveying emotions could be used to decode emotional valence from HGRs to facial expressions. Intracranial electroencephalography was recorded from 14 individuals while they participated in two tasks, one involving reading words with positive, negative, and neutral valence, and the other involving viewing faces with positive, negative, and neutral facial expressions. Quadratic discriminant analysis was used to identify information in the HGR that differentiates the three emotion conditions. A classifier was trained on the emotional valence labels from one task and was cross-validated on data from the same task (within-task classifier) as well as the other task (between-task classifier). Emotional valence could be decoded in the left medial orbitofrontal cortex and middle temporal gyrus, both using within-task classifiers and between-task classifiers. These observations suggest the presence of task-independent emotional valence information in the signals from these regions.

Key words: emotion; valence; intracranial EEG; classifier; decoding

Introduction

The ability to distinguish between negative, positive and neutral valence is a key part of emotion perception. In fact, one can scarcely define an emotional quality that is not either positive or negative in valence, as valence is an intrinsic characteristic of emotional experience and expression. A stimulus connoting negative valence suggests something aversive, unpleasant or repellent. It may lead one to exhibit defensive or self-protective reactions, to avoid further exposure and/or to experience unpleasant feelings, while a positively valenced stimulus may have the opposite effect. Humans have the ability to rapidly perceive valence from a wide variety of unrelated types of stimuli via virtually any sensory modality from the very basic (e.g. a noxious somatosensory stimulus) to the complex (e.g. a beautiful work of art), even when there is no consciously experienced feeling in response to the stimulus. The central conjecture evaluated in this study is that all instances of positive emotion and all instances of negative emotion are alike at some level that can be distinguished by the nervous system. In other words, we assess whether the neural circuit representation of emotional valence can be abstracted away from the actual stimulus and task features used to define the emotion concept.

A large body of research has been dedicated to identifying the neural mechanisms underlying emotion perception utilizing various methodologies. Invasive recordings from the human brain constitute a small proportion of this literature, but direct recordings from the human brain can overcome several limitations inherent in non-invasive technologies. Direct recording of neuronal activities allows for the measurement of brain responses with millisecond temporal resolution, millimeter spatial resolution and high signal-to-noise (SNR) ratio (Lachaux *et al.*, 2003). Invasive recordings can investigate deep brain areas not easily accessed with non-invasive electrophysiology. Intracranial electroencephalography (iEEG) studies of emotion perception have generally involved measuring event-related potentials or event-related spectral changes in response to emotionally laden and neutral stimuli (most commonly facial expressions, images of scenes, printed words or audio or video clips) and contrasting the

Received: 25 January 2021; **Revised:** 5 September 2021; **Accepted:** 22 December 2021

© The Author(s) 2021. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

responses between emotion conditions. A variety of limbic, paralimbic and frontal and temporal neocortical regions have been implicated (see Guillory and Bujarski (2014) for a review).

Emotion processing often recruits brain regions engaged in perception or interpretation of the stimulus, most commonly regions involved in visual processing as most tasks utilize visual stimuli (Vuilleumier and Driver, 2007; Boucher *et al.*, 2015; Weisholtz *et al.*, 2015). The involvement of sensory regions suggests that the neural substrates of emotion perception or processing are, to some degree, task specific. Nevertheless, several limbic and multimodal cortical regions have been implicated in emotion perception across various types of tasks. From the iEEG literature alone, such regions have included the amygdala with tasks involving viewing emotional scenes (Oya *et al.*, 2002), emotional facial expressions (Krolak-Salmon *et al.*, 2004; Pourtois *et al.*, 2010a,b; Sato *et al.*, 2011; Meletti *et al.*, 2012; Zheng *et al.*, 2017) or printed emotional words (Naccache *et al.*, 2005) and hearing vocal non-verbal emotional utterances (Dominguez-Borras *et al.*, 2019) or music (Omigie *et al.*, 2015); insula with tasks involving viewing emotional scenes (Brazdil *et al.*, 2009), facial expressions (Krolak-Salmon *et al.*, 2004) or printed emotional words (Ponz *et al.*, 2014); and orbitofrontal cortex in tasks involving viewing emotional facial expressions (Jung *et al.*, 2011) or emotion words (Ponz *et al.*, 2014) or listening to music (Omigie *et al.*, 2015).

These studies have examined neural responses to stimuli *within a particular task*, leaving open the question of the degree to which the emotion-related findings are specific to the particular task or reflect task-invariant emotion coding. We sought to identify brain regions coding for emotional valence *independent of processing domain* by comparing within-subject neural responses to similar valence defined in distinct ways. We considered visually presented stimuli with negative, neutral and positive valence from two separate tasks with different types of stimulus sets conveying emotion in different ways—one language based and the other image based. We focused on the HGR as this frequency band has shown correspondence with neural activation with good spatial and temporal resolution (Crone *et al.*, 2011; Lachaux *et al.*, 2012). One approach to identify task-invariant neural responses is to examine between-task decoding accuracy in a machine learning setting (Piva *et al.*, 2019). We trained machine learning classifiers to use the HGR to discriminate between the three emotion valence conditions in each task separately and identified brain regions in which classifier performance was better than chance for both tasks individually. The tasks differed in both the manner in which emotion was conveyed (facial expression or words) and in the specific type of emotion conveyed. Negative faces depicted expressions of fear, and positive faces depicted expressions of happiness, while the negative and positive words depicted a range of emotions related to depressive and counter-depressive themes. The two tasks were alike only in their valence categories (positive, neutral and negative). To assess the degree of task invariance, we further assessed the degree of extrapolation when the classifiers were trained on one task and tested on the other. This technique identified brain regions in which high-gamma signals contain information about emotional valence independent of the specific emotion conveyed or the method by which it is conveyed (words *vs* faces).

Materials and methods

Participants

Patients with pharmacologically intractable epilepsy who were undergoing intracranial EEG monitoring at New York Presbyterian Hospital, NYC, and at Brigham and Women's Hospital in Boston for seizure localization were recruited to participate after meeting the following inclusion criteria: they had capacity to consent, were fluent in English, were over 18 years old and were able to read. All protocols were approved by the IRB at each institution. The research was carried out in accordance with The Code of Ethics of the world Medical Association (Declaration of Helsinki) for experiments involving humans.

Tasks

Participants completed two similar tasks (one verbal and one non-verbal), involving the viewing of stimuli with positive, neutral and negative emotional valence that were presented on a laptop screen at the bedside. Stimuli were presented using either E-Prime (Psychology Software Tools, Inc.) or the Psychophysics Toolbox (Figure 1). The task was implemented in an identical way on each platform. Half of the participants completed the verbal task first and the other half completed the non-verbal task first. Most participants completed both tasks on the same day, but three participants completed them on consecutive days.

In the word (WD) task, stimuli consisted of single words presented in a white font within a white box on an otherwise black background, centered on the screen and subtending about 5–6° of visual angle vertically and 12–15° horizontally. There were 24 positive, 24 neutral and 24 negative words, which were either adjectives, nouns or verbs, chosen to be relevant to depressive and counter-depressive themes based on clinical experience and rated for suitability by a panel of three experts. Words were balanced across the categories for length, frequency within the lexicon and part of speech, with the exception that, within the neutral list, verbs were substituted for adjectives, given that adjectives are typically not free of emotional valence. Negative words included words such as *burden* and *guilty*. Positive words included words such as *praise* and *heroic*. Neutral words included words such as *clarinet* and *umbrella*. Example face stimuli are shown in Figures 1 and 2. The WD task was utilized in a previously published fMRI study (Epstein *et al.*, 2006). In the face (FA) task, participants viewed images from the NimStim Set, an image bank of validated emotional facial expressions (Tottenham *et al.*, 2009). Images consisted of color photographs of naturally posed actors of different sex and ethnicity from the neck up on a blank background exhibiting facial expressions of fear (negative condition), happiness (positive condition) or a blank expression (neutral condition). Images were centered on the screen and subtended approximately 17–20°

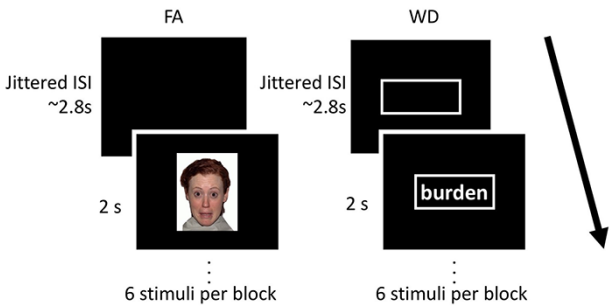

Fig. 1. Diagram of tasks. Each subject completed both a word (WD) task and a face (FA) task. Each task consisted of positive, neutral and negative stimuli with 24 trials per condition. Stimuli were presented in block design with six stimuli per block, 2 s presentation time and a jittered ISI around 2.8 s. Blocks were presented in pseudo-random order.

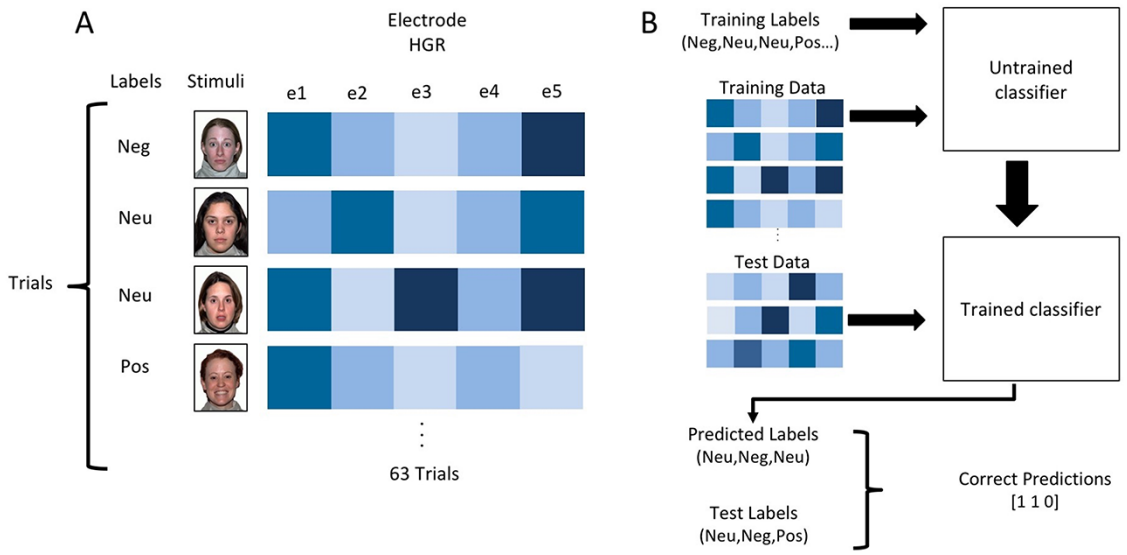

Fig. 2. HGRs from each of five electrodes in a particular brain region is entered as training data into a classifier along with condition labels for each trial. The trained classifier is then tested with data from three trials that were not included in the training set (one trial for each of the three conditions). This procedure is repeated until all trials have been tested, and the classifier performance is calculated as the percentage of trials correctly classified.

of visual angle vertically and 14–16° horizontally. There were 24 positive faces, 24 neutral faces and 24 negative faces.

In each task, the stimuli were presented one at a time in valence-specific six-stimulus blocks. Each stimulus appeared on the screen for 2 s followed by an inter-stimulus interval (ISI) jittered around an average of 2.8 s (range = 1.8–3.8 s). The participant was instructed to press a button with the right index finger in response to each stimulus, irrespective of the content. Participants were given up to 2 s to respond to each stimulus, and reaction time data were collected. Each task was analyzed separately by fitting the reaction times to a generalized linear mixed-effects model with condition (positive, negative or neutral) as a fixed effect and subject as the random effect and the natural logarithm as a link function.

Electrophysiology data collection

Electrodes consisted of commercially available strips, grids and depth electrodes that were implanted in various locations based on clinical need. The number, type and location of the electrodes was not influenced by the research plan and was dictated strictly by clinical needs. iEEG was recorded using the XLTEK clinical EEG recording system (Natus Neuroworks) with a sampling rate of 500 Hz for most participants. One study was sampled at 250 Hz, one was sampled at 2000 Hz and four studies were sampled at 512 Hz. The stimulus presentation laptop sent a trigger pulse to the EEG headbox that was recorded along with the EEG signals and was used to identify the precise timing of the stimulus presentation within the recordings.

Electrode localization

The iELVis software toolbox (Groppe *et al.*, 2017) was utilized to identify the precise locations of the intracranial electrodes. The Desikan–Killiany atlas (Desikan *et al.*, 2006), as implemented in iELVis and FreeSurfer (http://surfer.nmr.mgh.harvard.edu/), was used to label the locations of the cortical electrodes based on anatomical parcellation of each individual brain. Depth electrodes in hippocampus and amygdala were labeled based on FreeSurfer's volumetric brain segmentation (aparc + aseg.mgz).

Data analyses

Data analyses were carried out using MATLAB (Mathworks, Natick, MA). Electrodes were removed from the analyses if markedly corrupted by artifact, and line noise was removed by applying a series of notch filters at 60 Hz and harmonics. Each electrode was then re-referenced against the average signal. The high-gamma amplitude (HGA) was extracted by applying an 80–150 Hz bandpass filter on the re-referenced signals and then extracting the analytic signal from the Hilbert transform. Additional frequency bands were also tested and are described in Supplementary Methods and Supplementary Table S1.

HGR was calculated by subtracting the mean of the 1-s prestimulus baseline from the 1500 ms HGA signal post-stimulus onset. HGR was then binned into three 500 ms time windows representing the mean HGR during the first 500 ms following stimulus onset (bin 1), 500–1000 ms following stimulus onset (bin 2) and 1000–1500 ms following stimulus onset (bin 3).

Quadratic discriminant analysis was then used in order to identify information in the HGR that differentiates the three emotion conditions (Hung *et al.*, 2005; Meyers and Kreiman, 2012; Singer and Kreiman, 2012). A classifier (classify function in MATLAB Statistics and Machine Learning Toolbox) was trained separately for each brain region, task and time bin on the three different emotion conditions using a 'leave one out' cross-validation approach. To identify emotion-related information in the signal that is independent from the stimulus type, the classifier performance was also tested on the opposite task from which it was trained using a completely analogous procedure (we refer to this as 'between-task' classification, as opposed to 'within-task' classification when the classifier was tested on the same task on which it was trained).

Separate classifiers were trained and tested for each brain region containing at least five electrodes, combined across subjects. To reduce the impact of the multiple comparisons problem given the large number of classifiers and because our interest was in identifying brain regions that exhibited task-independent emotion information, we focused specifically on regions in which both within-task and between-task classifiers performed better

than chance. This combination was relatively unlikely to occur by chance, even with modest performance thresholds. *P*-values were computed for each region/time bin pair using the permutation method and corrected for multiple comparisons across the experiment. Region/time bin pairs were considered significant if the familywise error rate (FWER) was less than 0.05.

Among region-time bin pairs that survived the performance threshold, we investigated whether better than chance classifier performance was driven by the coding of valence or simply distinguishing emotion from non-emotion by comparing the proportion of emotion stimuli (positive and negative) that were classified with the correct valence as compared to the opposite valence using a one-sided binomial test. An analogous procedure was used to compare the proportion of emotion stimuli correctly classified *vs* misclassified as neutral, and among the misclassified emotion stimuli, the proportion labeled with the opposite valence as compared to the proportion misclassified as neutral.

Results

We recorded intracranial field potential signals in 14 participants (age 25–58, 6 female). The average reaction time across subjects was 986 ± 239 ms (mean ± SD, WD) and 871 ± 315 ms (FA). There was no significant effect of emotional valence for either the WD task ($P = 0.675$, ANOVA test) or the FA task ($P = 0.220$, ANOVA test). Reaction time was significantly shorter for faces than for words ($P < 0.001$).

Collectively, there were 947 intracerebral or subdural electrodes (Figure 3). High-gamma band (80–150 Hz) responses relative to pre-stimulus baseline were computed for each electrode. An example of HGR from an electrode in the left medial orbitofrontal cortex is shown in Figure 4 and Supplementary Figure S1. The neurophysiological responses from this electrode revealed a partial separation among the three emotional valences, particularly within the first second after stimulus onset, both for the FA task (Figure 4A) and for the WD task (Figure 4B). Notably, despite the large stimulus differences between the two tasks, the responses from this electrode were qualitatively similar between the two tasks: there was an increased HGR to negative (red) and neutral (black) stimuli compared to positive stimuli (green).

An ANOVA was performed to test whether HGR discriminated between the three valence conditions for each electrode, time bin and task. This involved 4290 statistical tests (3 time bins × 2 tasks × 715 electrodes). At a statistical threshold of $P < 0.05$, there were 222 significant tests (5.2% of the total), which is about what would be expected by chance. Because of the trial-to-trial variability in individual electrode responses, the small number of trials

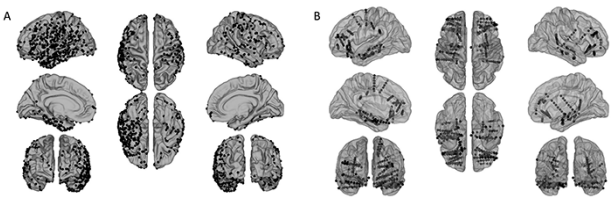

Fig. 3. Locations of all 947 electrodes transformed into standard coordinate space and plotted together on Freesurfer's average brain template. A. Surface electrodes. B. Depth electrodes (depicted with transparent cortical surfaces).

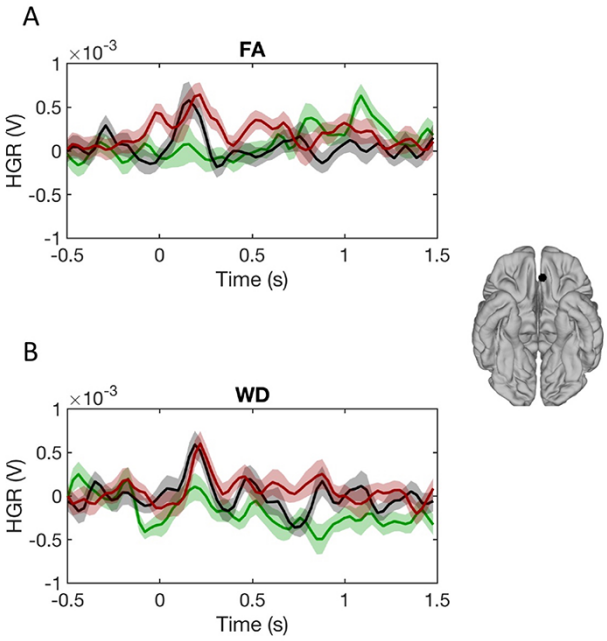

Fig. 4. Example electrode in the left mOFC showing high-gamma responses (DHG, normalized by the pre-stimulus baseline, Methods) in the face task (A) and word task (B). Responses are aligned to stimulus onset at Time = 0. Red = negative, black = neutral, green = positive. Shaded error bars indicate standard error of the mean ($n = 24$ trials). The location of the electrode is depicted on the freesurfer average brain template adjacent to the plots.

and the large number of electrodes, we used a classifier analysis based on ensembles of electrodes. Classifiers were trained to associate emotional valence labels for each trial with HGR data in three consecutive 500-ms time bins starting at stimulus onset. The procedure is able to determine in a data-driven way which electrodes and trials are most useful for classification. The classifiers were trained using cross-validation by randomly selecting a subset of the trials for a given emotional valence and task for training and testing its performance on the remaining trials (within-task classifier, Methods).

We examined each brain region containing at least five electrodes with an aim to identify brain regions in which HGR appeared sensitive to emotion independent of task in at least one of the three time bins. We defined this as better than chance classifier performance on both tasks individually and on at least one of the two between-task classifiers (training on words and testing on faces or vice versa). Among the 947 electrodes, 753 were localized to the amygdala, hippocampus or one of the cortical regions in the Desikan–Killiani atlas (most of the remaining electrodes were in white matter). Among these regions, there were 40 regions with at least five electrodes that were submitted for further analysis (Table 1). Because a language task was used, it was considered probable that some effects would be lateralized, and, thus, homologous regions from the two hemispheres were considered separately. Classifier accuracy (performance) was then tested on trials that were left out of the training set (Figure 2). The left medial orbitofrontal cortex (mOFC) during time bin 1 and the left middle temporal gyrus (MTG) during time bin 2 showed significantly better than chance classifier performance for both tasks individually and between tasks when trained on words and tested on faces for the high-gamma frequency band ($P < 0.005$; Figures 5 and 6; Supplementary Table S1 for findings

Table 1. Collective number of electrodes in each brain region

Region	N electrodes	Region	N electrodes	Region	N electrodes
Amygdala-L	**8**	**lateralorbitofrontal-L**	**20**	**rostralmiddlefrontal-L**	**30**
Amygdala-R	**7**	**lateralorbitofrontal-R**	8	**rostralmiddlefrontal-R**	**13**
Hippocampus-L	**6**	**lingual-L**	**10**	**superiorfrontal-L**	**13**
Hippocampus-R	**6**	lingual-R	1	**superiorfrontal-R**	**5**
bankssts-L	**9**	**medialorbitofrontal-L**	**6**	superiorparietal-L	3
bankssts-R	2	medialorbitofrontal-R	3	superiorparietal-R	4
caudalanteriorcingulate-L	2	**middletemporal-L**	**76**	**superiortemporal-L**	**66**
caudalanteriorcingulate-R	0	**middletemporal-R**	**16**	**superiortemporal-R**	**12**
caudalmiddlefrontal-L	**8**	**parahippocampal-L**	**10**	**supramarginal-L**	**39**
caudalmiddlefrontal-R	**6**	parahippocampal-R	4	**supramarginal-R**	**16**
entorhinal-L	**8**	**parsopercularis-L**	**20**	**temporalpole-L**	**10**
entorhinal-R	0	**parsopercularis-R**	**6**	temporalpole-R	2
frontalpole-L	0	**parsorbitalis-L**	**13**	transversetemporal-L	1
frontalpole-R	2	parsorbitalis-R	2	transversetemporal-R	0
fusiform-L	**28**	**parstriangularis-L**	**13**	cuneus-L	0
fusiform-R	**9**	**parstriangularis-R**	**6**	cuneus-R	0
inferiorparietal-L	**19**	**postcentral-L**	**42**	isthmuscingulate-L	0
inferiorparietal-R	**10**	**postcentral-R**	**13**	isthmuscingulate-R	0
inferiortemporal-L	**49**	**precentral-L**	**30**	paracentral-L	0
inferiortemporal-R	**16**	**precentral-R**	**18**	paracentral-R	0
insula-L	3	precuneus-L	1	pericalcarine-L	0
insula-R	2	precuneus-R	1	pericalcarine-R	0
lateraloccipital-L	**15**	rostralanteriorcingulate-L	1	posteriorcingulate-L	0
lateraloccipital-R	3	rostralanteriorcingulate-R	1	posteriorcingulate-R	0

Bolded regions contained ≥5 electrodes and were included in the analyses.

in other frequency bands). In both cases, the classifier trained on words performed better than chance when tested on both words and faces. The classifier trained on faces performed better than chance when tested on faces but did not exceed threshold when tested on words. Mean linear coefficients are depicted in Supplementary Figure S2. Because these two regions contained markedly different numbers of electrodes (6 in the left mOFC, 76 in the left MTG), the left MTG was re-analyzed for time bin 2 (500–1000 ms) using different random subsamples of six electrodes from this region for each iteration of the classifier. With electrode subsampling, the findings were no longer significant in this region, suggesting that there are subsets of electrodes that drive the classification performance. Electrode weights, as estimated from the absolute value of the mean linear coefficients, are depicted by location in Supplementary Figure S3.

As the classifier labeled trials from among three categories, better than chance performance could be achieved even if only one of the three categories could be discriminated from the other two. Neutral stimuli lack emotional content and are qualitatively different from the other two categories for this reason. Thus, we explored whether the classifier's success in mOFC and MTG depended only on an ability to discriminate emotion from no emotion or whether positive stimuli could be correctly discriminated from negative stimuli. We found that across the four classifier analyses (the two within-task analyses and the two between-task analyses), the emotion stimuli (positive and negative trials) were correctly labeled more often than they were labeled with the opposite emotional valence, both in the left mOFC during bin 1 (81 correct emotion labels (46% of emotion trials), 39 incorrect emotion labels (22%), $P < 0.0001$) and in the left MTG during bin 2 (75 correct emotion labels (43%), 50 incorrect emotion labels (28%), $P = 0.016$; Figure 7; see Supplementary Figure S4 for full confusion matrices).

Fig. 5. Electrode locations in the left mOFC (red) and left MTG (brown). Brain regions colored dark gray were included in the analyses but did not show significant results. Brain regions colored light gray were excluded from analysis due to inadequate electrode coverage (<5 electrodes).

Discussion

Social interactions constitute the essential fabric of daily experience. Social interactions depend on each individual's ability to recognize emotions expressed by others either verbally or nonverbally. Here, we sought to identify neural substrates of emotional valence and to assess whether those neural substrates represent abstract emotional concepts or task-specific signals. Consistent with earlier work, we found that neural responses could distinguish between different emotional valences (Figure 2) both in a task involving language and a task involving faces (Figure 1). We used a machine learning classifier to quantify the extent to which emotional valence could be read out in single

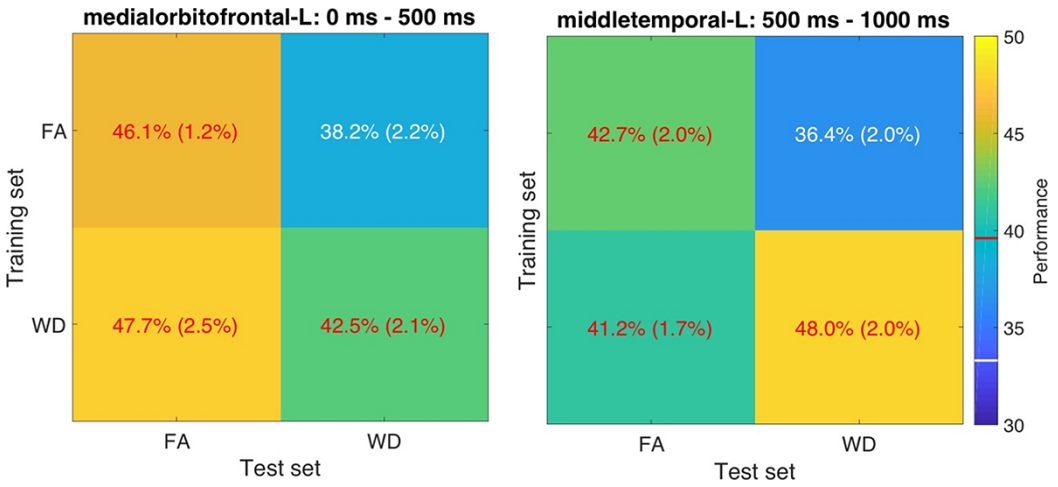

Fig. 6. Within-task and cross-task mean classifier performance (standard deviations in parentheses) for the L mOFC in bin 1 and the L MTG in bin 2. These two regions showed better than chance performance for the within-task classifiers for both words and faces as well as one of the cross-task classifiers. Performance colored red indicates it exceeds the significance threshold for $P < 0.05$. On the color bar, the white line indicates chance performance (33.3%), and the red line indicates the threshold for performance significantly better than change ($P < 0.05$). NOTE: This figure requires color.

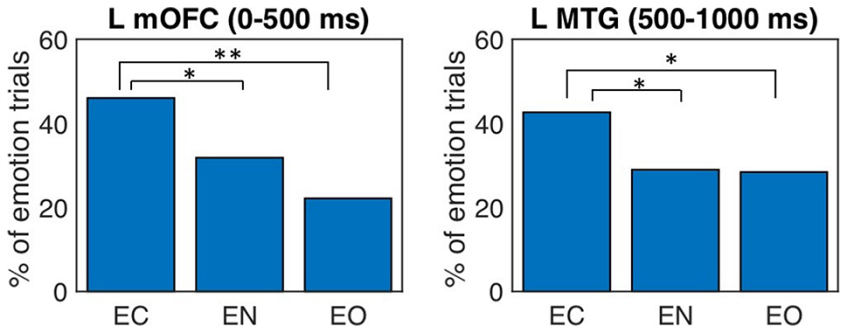

Fig. 7. To evaluate whether the classifier's success in mOFC and MTG depended only on an ability to discriminate emotion from no emotion (neutral stimuli) or whether positive stimuli could be correctly discriminated from negative stimuli, we examined the misclassification pattern among emotion stimuli (positive and negative faces and words). Combined across the four classifiers (FA-FA, FA-WD, WD-FA, WD-WD), emotion stimuli were more likely to be classified with the correct valence (EC) than with the opposite valence (EO) in the L mOFC and L MTG, indicating that the signal contained information discriminating the two emotional valences from each other. Emotion stimuli were also more likely to be classified correctly than misclassified as neutral. EC = emotion stimuli classified correctly; EN = emotion stimuli misclassified as neutral; EO = emotion stimuli misclassified as the opposite emotion valence; * = $P < 0.05$; ** = $P < 0.0001$.

trials (Figure 3). The classifier was able to discriminate emotional valence when trained and tested on different partitions of the trials *within* each task, consistent with a body of earlier work demonstrating task-specific representation of emotional valence throughout multiple brain regions. Two brain regions, the MTG and OFC stood out from the rest because their representation allowed the classifier to extrapolate *between* tasks (Figure 5).

The sequence of cortical activation involved in the processing of a stimulus generally follows a pathway beginning in primary sensory cortex and propagating to higher cortical areas with prominent feedback modulation at multiple stages of processing as features of the stimulus are decoded. Different types of emotional stimuli may require distinct processing steps to decode the emotional valence. For example, within the visual modality, some investigators have suggested that the analysis of low-frequency visual features of fearful facial expressions may be adequate to activate the amygdala via a magnocellular retinal-collicular-pulvinar pathway that bypasses visual cortex (Vuilleumier *et al.*, 2003). In contrast, representing the emotional content in printed words requires fine-grained decoding of high spatial-frequency information to represent the visual word form, and lexicosemantic transformation to decode word meaning that likely involves peri-Sylvian language areas (Weisholtz *et al.*, 2015). Emotional content in stimuli can modulate activity at multiple stages of processing specific to a particular task, including areas of language cortex (Beauregard *et al.*, 1997; Maddock *et al.*, 2003; Cato *et al.*, 2004; Kuchinke *et al.*, 2005), visual cortex (Vuilleumier *et al.*, 2001; Pessoa *et al.*, 2002) and auditory cortex (Sander and Scheich, 2001; Grandjean *et al.*, 2005; Liebenthal *et al.*, 2016). It is clear that emotion impacts the perceptual/cognitive processing stream in a manner that is to some extent dependent on the particularities of the stimuli used to convey the emotion. While these neural changes can be utilized by a machine learning classifier to decode emotional valence categories, it is unclear if emotion is truly coded in these perceptual/cognitive areas or if the neural changes reflect augmentation of perceptual/cognitive processing of the emotional stimuli.

At a basic level, emotional valence has meaning *independent* of task or stimulus type. Classifier decoding analysis can be used to identify valence-related information in a neural signal that is

independent of the particular task or stimulus type if a classifier trained on one task is able to decode the stimulus valence from a qualitatively different stimulus set. There was little similarity in the sensory inputs between the verbal stimuli in the WD task and the face images in the FA task, aside from the fact that the emotional valence of the stimuli could be broadly categorized as negative, neutral and positive. Nevertheless, despite the heterogeneity between the two tasks, we found that some classifiers could not only read out valence information within a task but also *across tasks*. Specifically, within the mOFC and MTG, classifiers trained exclusively within the WD task were able to extract valence information when considering neural responses during the FA task on which they were not trained. This indicates that the mOFC and MTG may represent emotional valence-related information independent of the representation of the particular stimuli or the manner in which that emotional content is discerned from the stimuli (rapid visual detection *vs* lexicosemantic transformation). Thus, the between-task extrapolation effect is likely not driven simply by emotional modulation of circuitry involved in processing facial identity or word meaning. The lack of symmetry between the two between-task classifiers was interesting but not necessarily surprising. Our criteria for identifying task-invariant valence information required better than chance decoding within each task as well as with at least one of the two between-task classifiers. It was not expected that just because training on one task allowed for successful decoding in the other task that the converse must also be true. One possible explanation for the asymmetry is differing SNR ratios between the two tasks. A classifier trained on a task with higher SNR may perform better when tested on a lower SNR dataset than the converse.

Both the mOFC and the MTG are high-order multimodal cortical association areas that have been implicated in emotional processing. The OFC has been closely linked with processing of emotion-related information supporting goal-directed behavior. It has been proposed that OFC represents changing and relative reward values (Kringelbach and Rolls, 2004) and that it may represent the reward and punishment value of primary as well as learned reinforcers, allowing for behavior change to occur when reinforcement values change (Rolls, 2000). Thus, the OFC appears to monitor the affective properties of stimuli from various sensory modalities and is therefore ideally situated to process valence in a task-invariant manner. The OFC has been implicated in the processing of both emotional facial expressions and emotion words. Ventral frontal lobe damage can lead to impairment in identification of facial expressions even in patients who were not impaired in facial recognition (Hornak *et al.*, 1996). Bilateral OFC lesions can also cause impairment in emotional voice discrimination (Hornak *et al.*, 2003). Magnetoencephalography can detect early involvement of the OFC in processing affectively charged visual scenes (Rudrauf *et al.*, 2008) and phase-locking between the OFC and amygdala in response to emotional facial expressions (Cushing *et al.*, 2019). The emotional valence of written words can also modulate OFC activation seen with functional MRI (Lewis *et al.*, 2007).

The middle temporal gyrus is a multimodal association area on the lateral temporal lobe bounded inferiorly and posteriorly by visual association cortex and superiorly by auditory association cortex. Lesions to this region can lead to deficits in word comprehension and naming (Dronkers *et al.*, 2004) and its functional and structural connectivity with peri-Sylvian language areas position it as an important region for language comprehension (Turken and Dronkers, 2011) and possibly semantic processing more generally (Binder and Desai, 2011). Functional imaging studies have shown that emotional content in words can modulate MTG activity (Beauregard *et al.*, 1997; Cato *et al.*, 2004; Weisholtz *et al.*, 2015). The posterior portion of the MTG and adjacent superior temporal sulcus (STS) are also involved in face processing and have been implicated, in particular, in perception of facial expression (Haxby *et al.*, 2000, 2002; Said *et al.*, 2011). It has been proposed that the posterior MTG/STS represents changeable aspects of faces independent of facial identity (Haxby *et al.*, 2000), although the notion of two truly dissociable systems for the recognition of facial identity and facial expression has been questioned (Calder and Young, 2005). In a human iEEG study, a decoding analysis was able to discriminate fearful and happy facial expressions using information from the high-gamma band and below 30 Hz in the lateral and inferior temporal cortex, although performance was better in the inferior temporal cortex, contrary to prediction (Tsuchiya *et al.*, 2008). Emotional scenes (Sabatinelli *et al.*, 2011) and emotional gestures (Grosbras and Paus, 2006; Flaisch *et al.*, 2009) have also been shown to modulate activity in portions of lateral temporal neocortex. The variety of types of emotional stimuli that engage the MTG may indicate an emotion function independent of stimulus type or task, but it is also possible that emotion modulates various types of stimulus representations in the MTG and adjacent regions. The fact that the MTG classifier could decode facial expressions when trained on word valence may indicate regions of MTG that can represent emotional valence more generally. Alternatively, emotional valence may modulate representations of words and faces that have overlapping anatomical fields, at least within the spatial resolution of an iEEG electrode.

Variability in the neural response to stimuli in the same category (with the same emotional valence label) can occur due to noise in the signal, differences in degree to which different stimuli evoke the emotional connotations they are intended to evoke and distractions during the task. Typically, such variability is dealt with by averaging across trials, which assigns equal weight to each trial. This approach risks missing the signal within the noise when there is a small number of trials. The classifier analysis allows decoding at the single trial level and is sensitive to relevant information in the signal, even with a small number of trials, as trials (and electrodes) containing information relevant to the condition labels can be weighted more strongly than those that do not. Similarly, responses may vary from electrode-to-electrode within a brain region due to a variety of factors, such as electrode artifact, epileptiform activity, or anatomic distributions that do not map properly onto the gyral patterns reflected in the Desikan–Killiany atlas, and averaging across electrodes within a region may obscure findings by assigning equal weight to relevant and irrelevant electrodes. Decoding analysis uses a data-driven approach to assign weight to the most informative electrodes and trials at the expense of some loss of temporal and spatial precision. The MTG is a considerably larger region than the mOFC, and in our study, there were considerably more electrodes covering the left MTG than the left mOFC (76 *vs* 6). We repeated the analyses of the left MTG randomly subsampling the electrodes down to 6 with each iteration of the classifier so as to equalize the amount of data and make the performance results more comparable between the two regions. However, this resulted in the MTG classifiers no longer performing better than chance. While the MTG was fairly well-covered by electrodes (Figure 5), the region is functionally heterogenous, and it is likely that all electrodes did not contribute equally to the classifier performance. Randomly sampling only 6 out of the 76 electrodes likely did not consistently include enough relevant electrodes to mirror the performance of the classifiers that included all 76 electrodes.

The classifier was trained to distinguish three different valence categories, but the classifier performances we report could have been achieved even if the classifier was only able to distinguish one of the stimulus types from the other two. For example, if a signal distinguishes neutral stimuli from emotion-laden stimuli but represents positive and negative valence similarly, a classifier might decode neutral stimuli very successfully but could achieve, at best, 50% performance decoding the positive and neutral stimuli. In this scenario, a classifier could achieve, in principle, an overall performance level as high as 67%. We investigated this possibility and found that within the left mOFC and left MTG, emotion trials were more likely to be labeled with the correct emotional valence than the incorrect emotional valence, suggesting that positive and negative emotions can be discriminated from each other in these regions. Thus, the neural signal contains information about emotional valence and not just the presence or absence of emotional content.

Limitations of this study included the low number of trials per condition and variable electrode locations across participants. As with any study employing invasive human brain recordings, the participants are limited to a clinical population (in this case, patients with epilepsy) in whom neurophysiological properties can differ from healthy individuals. The low number of trials likely contributed to unconvincing findings at the single electrode level. While the classifier analysis allowed for the identification of task-invariant emotional valence encoding, the need to bin signals across time and combine electrodes within brain regions limited the spatial and temporal specificity of the findings. HGA was used as a metric of brain activity based on a body of evidence demonstrating consistent and well-localized task-related activation of sensorimotor and language areas, but additional valence-related information is likely encoded in other frequency bands as well (Supplementary Table S1). The amygdala is known to be involved in representing emotional properties of experimental stimuli but did not appear as a significant finding in the primary analysis of this study. It is possible that with a greater number of trials or amygdala electrodes, such an effect may have been detected, but it also may be that amygdala activity contains more valence-relevant information at other frequency bands. In fact, when other frequency bands were examined in a secondary analysis, the right amygdala showed significant task-invariant valence information in the low gamma band (30–80 Hz) during bin 1 ($P = 0.035$; Supplementary Table S1). The between-task classifier appears to be a promising approach for the identification of task-invariant information in neural signals, but further research is needed to clarify the temporal dynamics of these signals as well as the spatial specificity. Additionally, different parts of the brain may carry information in different frequency bands, and further research is needed to understand the relationships between frequency band, brain location, and task.

Conclusions

Viewing negatively valenced, positively valenced and neutral stimuli evoked changes in the high-gamma band that differentiated between the three valence conditions in the left mOFC and left MTG. The signal in these regions contains valence-related information that is independent of the method by which the emotional valence is conveyed (e.g. via facial expression or words) by showing that a classifier trained to decode emotion from words can perform better than chance when decoding emotion from facial expressions, even when it has not been trained on facial expression data at all. The results suggest that mOFC and MTG encode general stimulus-independent valence-related information that can be applied in different contexts and may provide a mechanism by which qualitatively different items can be compared based on emotional valence.

Funding

This work was supported by NIH (MH107820, D.W. and G.K.; EY026025, G.K.) and the Epilepsy Foundation (T.B.).

Conflict of interest

None declared.

Supplementary data

Supplementary data is available at *SCAN* online.

References

Beauregard, M., Chertkow, H., Bub, D., Murtha, S., Dixon, R., Evans, A. (1997). The neural substrate for concrete, abstract, and emotional word lexica a positron emission tomography study. *Journal of Cognitive Neuroscience*, **9**(4), 441–61.

Binder, J.R., Desai, R.H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences*, **15**(11), 527–36.

Boucher, O., D'Hondt, F., Tremblay, J., *et al.* (2015). Spatiotemporal dynamics of affective picture processing revealed by intracranial high-gamma modulations. *Human Brain Mapping*, **36**(1), 16–28.

Brazdil, M., Roman, R., Urbanek, T., *et al.* (2009). Neural correlates of affective picture processing—a depth ERP study. *NeuroImage*, **47**(1), 376–83.

Calder, A.J., Young, A.W. (2005). Understanding the recognition of facial identity and facial expression. *Nature Reviews Neuroscience*, **6**(8), 641–51.

Cato, M.A., Crosson, B., Gokcay, D., *et al.* (2004). Processing words with emotional connotation: an FMRI study of time course and laterality in rostral frontal and retrosplenial cortices. *Journal of Cognitive Neuroscience*, **16**(2), 167–77.

Crone, N.E., Korzeniewska, A., Franaszczuk, P.J. (2011). Cortical gamma responses: searching high and low. *International Journal of Psychophysiology*, **79**(1), 9–15.

Cushing, C.A., Im, H.Y., Adams, R.B., Jr, Ward, N., Kveraga, K. (2019). Magnocellular and parvocellular pathway contributions to facial threat cue processing. *Social Cognitive and Affective Neuroscience*, **14**(2), 151–62.

Desikan, R.S., Segonne, F., Fischl, B., *et al.* (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, **31**(3), 968–80.

Dominguez-Borras, J., Guex, R., Mendez-Bertolo, C., *et al.* (2019). Human amygdala response to unisensory and multisensory emotion input: no evidence for superadditivity from intracranial recordings. *Neuropsychologia*, **131**, 9–24.

Dronkers, N.F., Wilkins, D.P., Van Valin, R.D., Jr, Redfern, B.B., Jaeger, J.J. (2004). Lesion analysis of the brain areas involved in language comprehension. *Cognition*, **92**(1–2), 145–77.

Epstein, J., Pan, H., Kocsis, J.H., *et al.* (2006). Lack of ventral striatal response to positive stimuli in depressed versus normal subjects. *American Journal of Psychiatry*, **163**(10), 1784–90.

Flaisch, T., Schupp, H.T., Renner, B., Junghofer, M. (2009). Neural systems of visual attention responding to emotional gestures. *NeuroImage*, **45**(4), 1339–46.

Grandjean, D., Sander, D., Pourtois, G., *et al.* (2005). The voices of wrath: brain responses to angry prosody in meaningless speech. *Nature Neuroscience*, **8**(2), 145–6.

Groppe, D.M., Bickel, S., Dykstra, A.R., *et al.* (2017). iELVis: an open source MATLAB toolbox for localizing and visualizing human intracranial electrode data. *Journal of Neuroscience Methods*, **281**, 40–8.

Grosbras, M.H., Paus, T. (2006). Brain networks involved in viewing angry hands or faces. *Cerebral Cortex*, **16**(8), 1087–96.

Guillory, S.A., Bujarski, K.A. (2014). Exploring emotions using invasive methods: review of 60 years of human intracranial electrophysiology. *Social Cognitive and Affective Neuroscience*, **9**(12), 1880–9.

Haxby, J.V., Hoffman, E.A., Gobbini, M.I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, **4**(6), 223–33.

Haxby, J.V., Hoffman, E.A., Gobbini, M.I. (2002). Human neural systems for face recognition and social communication. *Biological Psychiatry*, **51**(1), 59–67.

Hornak, J., Rolls, E.T., Wade, D. (1996). Face and voice expression identification in patients with emotional and behavioural changes following ventral frontal lobe damage. *Neuropsychologia*, **34**(4), 247–61.

Hornak, J., Bramham, J., Rolls, E.T., *et al.* (2003). Changes in emotion after circumscribed surgical lesions of the orbitofrontal and cingulate cortices. *Brain*, **126**(Pt 7), 1691–712.

Hung, C.P., Kreiman, G., Poggio, T., DiCarlo, J.J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science*, **310**(5749), 863–6.

Jung, J., Bayle, D., Jerbi, K., *et al.* (2011). Intracerebral gamma modulations reveal interaction between emotional processing and action outcome evaluation in the human orbitofrontal cortex. *International Journal of Psychophysiology*, **79**(1), 64–72.

Kringelbach, M.L., Rolls, E.T. (2004). The functional neuroanatomy of the human orbitofrontal cortex: evidence from neuroimaging and neuropsychology. *Progress in Neurobiology*, **72**(5), 341–72.

Krolak-Salmon, P., Henaff, M.A., Vighetto, A., Bertrand, O., Mauguiere, F. (2004). Early amygdala reaction to fear spreading in occipital, temporal, and frontal cortex: a depth electrode ERP study in human. *Neuron*, **42**(4), 665–76.

Kuchinke, L., Jacobs, A.M., Grubich, C., Vo, M.L., Conrad, M., Herrmann, M. (2005). Incidental effects of emotional valence in single word processing: an fMRI study. *NeuroImage*, **28**(4), 1022–32.

Lachaux, J.P., Rudrauf, D., Kahane, P. (2003). Intracranial EEG and human brain mapping. *Journal of Physiology-Paris*, **97**(4–6), 613–28.

Lachaux, J.P., Axmacher, N., Mormann, F., Halgren, E., Crone, N.E. (2012). High-frequency neural activity and human cognition: past, present and possible future of intracranial EEG research. *Progress in Neurobiology*, **98**(3), 279–301.

Lewis, P.A., Critchley, H.D., Rotshtein, P., Dolan, R.J. (2007). Neural correlates of processing valence and arousal in affective words. *Cerebral Cortex*, **17**(3), 742–8.

Liebenthal, E., Silbersweig, D.A., Stern, E. (2016). The language, tone and prosody of emotions: neural substrates and dynamics of spoken-word emotion perception. *Frontiers in Neuroscience*, **10**, 506.

Maddock, R.J., Garrett, A.S., Buonocore, M.H. (2003). Posterior cingulate cortex activation by emotional words: fMRI evidence from a valence decision task. *Human Brain Mapping*, **18**(1), 30–41.

Meletti, S., Cantalupo, G., Benuzzi, F., *et al.* (2012). Fear and happiness in the eyes: an intra-cerebral event-related potential study from the human amygdala. *Neuropsychologia*, **50**(1), 44–54.

Meyers, E., Kreiman, G. (2012). Tutorial on pattern classification in cell recording. In: Kriegeskorte, N., Kreiman, G., editors. *Visual Population Codes: Toward a Common Multivariate Framework for Cell Recording and Functional Imaging*, Cambridge, MA: MIT Press. 517–38.

Naccache, L., Gaillard, R., Adam, C., *et al.* (2005). A direct intracranial record of emotions evoked by subliminal words. *Proceedings of the National Academy of Sciences*, **102**(21), 7713–7.

Omigie, D., Dellacherie, D., Hasboun, D., *et al.* (2015). An intracranial EEG study of the neural dynamics of musical valence processing. *Cerebral Cortex*, **25**(11), 4038–47.

Oya, H., Kawasaki, H., Howard, M.A., 3rd, Adolphs, R. (2002). Electrophysiological responses in the human amygdala discriminate emotion categories of complex visual stimuli. *The Journal of Neuroscience*, **22**(21), 9502–12.

Pessoa, L., McKenna, M., Gutierrez, E., Ungerleider, L.G. (2002). Neural processing of emotional faces requires attention. *Proceedings of the National Academy of Sciences*, **99**(17), 11458–63.

Piva, M., Velnoskey, K., Jia, R., Nair, A., Levy, I., Chang, S.W. (2019). The dorsomedial prefrontal cortex computes task-invariant relative subjective value for self and other. *Elife*, **8**, e44939.

Ponz, A., Montant, M., Liegeois-Chauvel, C., *et al.* (2014). Emotion processing in words: a test of the neural re-use hypothesis using surface and intracranial EEG. *Social Cognitive and Affective Neuroscience*, **9**(5), 619–27.

Pourtois, G., Spinelli, L., Seeck, M., Vuilleumier, P. (2010a). Modulation of face processing by emotional expression and gaze direction during intracranial recordings in right fusiform cortex. *Journal of Cognitive Neuroscience*, **22**(9), 2086–107.

Pourtois, G., Spinelli, L., Seeck, M., Vuilleumier, P. (2010b). Temporal precedence of emotion over attention modulations in the lateral amygdala: intracranial ERP evidence from a patient with temporal lobe epilepsy. *Cognitive, Affective and Behavioral Neuroscience*, **10**(1), 83–93.

Rolls, E.T. (2000). The orbitofrontal cortex and reward. *Cerebral Cortex*, **10**(3), 284–94.

Rudrauf, D., David, O., Lachaux, J.P., *et al.* (2008). Rapid interactions between the ventral visual stream and emotion-related structures rely on a two-pathway architecture. *Journal of Neuroscience*, **28**(11), 2793–803.

Sabatinelli, D., Fortune, E.E., Li, Q., *et al.* (2011). Emotional perception: meta-analyses of face and natural scene processing. *NeuroImage*, **54**(3), 2524–33.

Said, C.P., Haxby, J.V., Todorov, A. (2011). Brain systems for assessing the affective value of faces. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **366**(1571), 1660–70.

Sander, K., Scheich, H. (2001). Auditory perception of laughing and crying activates human amygdala regardless of attentional state. *Cognitive Brain Research*, **12**(2), 181–98.

Sato, W., Kochiyama, T., Uono, S., *et al.* (2011). Rapid amygdala gamma oscillations in response to fearful facial expressions. *Neuropsychologia*, **49**(4), 612–7.

Singer, J., Kreiman, G. (2012). Introduction to statistical learning and pattern classification. In: Kriegeskorte, N., Kreiman, G., editors. *Visual Population Codes: Toward a Common Multivariate Framework for Cell Recording and Functional Imaging*, Cambridge, MA: MIT Press. 497–516.

Tottenham, N., Tanaka, J.W., Leon, A.C., *et al.* (2009). The NimStim set of facial expressions: judgments from untrained research participants. *Psychiatry Research*, **168**(3), 242–9.

Tsuchiya, N., Kawasaki, H., Oya, H., Howard, M.A., 3rd, Adolphs, R. (2008). Decoding face information in time, frequency and space

from direct intracranial recordings of the human brain. *PLoS One*, **3**(12), e3892.

Turken, A.U., Dronkers, N.F. (2011). The neural architecture of the language comprehension network: converging evidence from lesion and connectivity analyses. *Frontiers in System Neuroscience*, **5**, 1.

Vuilleumier, P., Armony, J.L., Driver, J., Dolan, R.J. (2001). Effects of attention and emotion on face processing in the human brain: an event-related fMRI study. *Neuron*, **30**(3), 829–41.

Vuilleumier, P., Armony, J.L., Driver, J., Dolan, R.J. (2003). Distinct spatial frequency sensitivities for processing faces and emotional expressions. *Nature Neuroscience*, **6**(6), 624–31.

Vuilleumier, P., Driver, J. (2007). Modulation of visual processing by attention and emotion: windows on causal interactions between human brain regions. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **362**(1481), 837–55.

Weisholtz, D.S., Root, J.C., Butler, T., *et al.* (2015). Beyond the amygdala: linguistic threat modulates peri-sylvian semantic access cortices. *Brain and Language*, **151**, 12–22.

Zheng, J., Anderson, K.L., Leal, S.L., *et al.* (2017). Amygdala-hippocampal dynamics during salient information processing. *Nature Communications*, **8**, 14413.