Sensory encoding of emotion conveyed by the face and visual context

Katherine Soderberg[1]
Grace Jang[1]
Philip Kragel[1]

[1]Emory University, Department of Psychology

* Please address correspondence to:

Philip A. Kragel
Department of Psychology, PAIS 475
Emory University
Atlanta, GA 30032
404-727-3409

Abstract

Humans rapidly detect and interpret sensory signals that have emotional meaning. Facial expressions—particularly important signifiers of emotion—are processed by a network of brain regions including amygdala and posterior superior temporal sulcus (pSTS). However, the precise computations these regions perform, and whether emotion-specific representations explain their responses to socially complex, dynamic stimuli remain contentious. Here we investigated whether representations from artificial neural networks (ANNs) optimized to recognize emotion from facial expressions alone or the broader visual context differ in their ability to predict human pSTS and amygdala activity. We found that representations of facial expressions were encoded in pSTS, but not the amygdala, whereas representations related to visual context were encoded in both regions. These findings demonstrate how the pSTS may operate on abstract representations of facial expressions, such as ‘fear’ and ‘joy’, whereas the amygdala encodes the emotional significance of visual information more broadly.

Introduction

Humans are constantly confronted with sensory signals that have emotional meaning. A grimace, slumped shoulders, the arrangement of people in a scene—all of these sensory events convey affective information critical for navigating a dynamic environment. Though these signals are varied and complex, humans rapidly detect and interpret them. A particularly salient source of signals is facial expressions, which convey information about emotional states. Perceiving the emotion expressed by the face, humans can infer an individual's intentions and motivations, which are crucial for successful social interaction. Along with facial expressions, other elements of the visual array including actions, objects, and scenes can signify emotional meaning. Detecting the affective significance of one's present environment powerfully motivates and constrains behavior. How does the brain accomplish this sensory encoding of emotion? Studies of face perception and emotion processing implicate the superior temporal sulcus (STS) and amygdala as important nodes in distributed neural networks for registering emotional information in the sensory environment.

Research in the field of cognitive neuroscience has revealed a distributed network of regions that process faces, extracting representations of who the face belongs to and the emotion it conveys. Bruce and Young's influential model posited that cognitive processing of faces occurs in independent streams for identity and emotional expression[1]. Haxby and colleagues extended this idea, suggesting that facial features are processed by hierarchically organized systems for face processing[2]. In this and related models of face processing[3,4], dorsal visual regions process changeable facial features and integrate them with other signals, while regions in the ventral visual stream process static features. The posterior STS is thought to integrate dynamic information from facial expressions, along with other social signals (e.g., body posture and vocal

tone), and map expressions to emotion categories[5–7]. The amygdala, on the other hand, plays a role in extended face processing systems by sensing threat and other salient affective information from facial expressions and has been shown to discriminate between fearful and non-fearful faces[8–10]. Thus, the posterior STS and the amygdala have been robustly implicated in processing emotional faces, each with dissociable functions.

The posterior STS responds to a variety of sensory signals that guide the perception of facial expressions. Broadly, it is sensitive to biological motion, from minimal arrays to complex action sequences[11,12]. It is consistently engaged by facial expressions and its responses to these expressions are sufficient to decode emotion category[7,13]. Furthermore, the posterior STS responds to emotional cues originating from other sources, including tone of voice and body posture, which also contain information from which emotion category can be decoded[14,15]. These results have led some to suggest that the posterior STS encodes supramodal, or amodal representations of emotion[16]. Such representations, which consist of emotion concepts abstracted away from their inputs, result from processing of modality-specific features from distinct sensory pathways that are mapped to more abstract emotion categories.

The amygdala, a component of extended face processing systems, has been associated with the detection of salient exteroceptive events, including innate and learned threats[17,18], rewards[19], animate objects[20], and multiple different facial expressions[21]. Evidence is mixed as to whether the amygdala responds preferentially to one or several facial expressions as opposed to facial emotion more generally. A long line of studies shows that the amygdala preferentially responds to fearful (and sometimes angry and disgusted) expressions compared to neutral ones[22–24]. More recent work using multivariate decoding approaches found that amygdala activity could differentiate fearful from non-fearful faces, but not other categories[10]. However, there is evidence

against specificity for fear, as amygdala activity has been found to reflect the intensity or ambiguity of facial expressions rather than a single category such as fear[25,26]. Given its broader role in processing both threats and rewards, amygdala responses to facial expressions of fear and anger have most commonly been interpreted as involving the detection of threat[27] or salience more broadly[28,29]. Debate continues over whether amygdala responses to expressions of fear reflect holistic encoding of that emotion category, or whether this activity reflects representations related to broader dimensions of salience, valence, or intensity rather than representations of facial expressions themselves.

Given this evidence, it is clear that the amygdala and posterior STS are involved in processing emotional facial expressions. However, much of our understanding of these two regions is based on research using paradigms with decontextualized facial expressions, such that the face dominates the visual array. For example, a prominent paradigm shows subjects cropped fearful and angry faces, devoid of any context[30]. In this and similar tasks, observing greater activity in the amygdala or posterior STS could mean that these regions encode representations of specific facial expressions, or that they encode representations related to the broader visual context that are not specific to facial expressions *per se*. Accordingly, it is an open question as to the generalizability of representations encoded in these regions; it is possible that the information encoded by either or both of these regions is better explained by visual context as opposed to facial expressions specifically.

Here we evaluate the nature of representations encoded in the amygdala and posterior STS using artificial neural networks (ANNs) as models of human brain systems. This approach is part of a broader framework that aims to identify the neural computations that enable complex behaviors such as object and face recognition[31]. Modeling efforts using deep convolutional

networks to approximate the ventral visual stream have provided multiple insights about object recognition[32] and face processing[33–35], including functional specialization in the fusiform gyrus[36]. However, they have been less successful at explaining activity in dorsal and lateral pathways[37], and emotion recognition in particular. Further, this work has largely characterized brain responses to tightly controlled experimental paradigms that do not mimic experiences from everyday life because they do not present faces in rich social contexts with concurrent multimodal stimuli[38]. We aimed to address these issues by testing whether ANNs with different architectures and training objectives can better explain the functions of the posterior STS and amygdala to naturalistic videos. We compare the performance of a deep convolutional network with attention mechanisms trained to recognize facial expressions and predict the location of specific facial landmarks (EmoFAN)[39] and a deep convolutional network trained to classify the emotion schema present in visual scenes (EmoNet)[40]. Using features extracted by these ANNs to develop encoding models that predict brain activity in a given region[41], we characterize how the amygdala and posterior STS disentangle the emotional meaning of facial expressions and the broader visual context.

Results

To investigate face processing in a rich naturalistic context, we examined human brain activity measured as subjects ($N = 20$) watched the full-length movie *500 Days of Summer* while undergoing fMRI (~1.5 hours per participant; data sampled from the Naturalistic Neuroimaging Database[42]). Using both deep convolutional neural networks, we extracted features from each timepoint in the movie and fit encoding models to predict brain activity in the amygdala and posterior STS (Figure 1). This enabled us to test different accounts of emotion processing (expression-specific vs. general visual context) for each region.

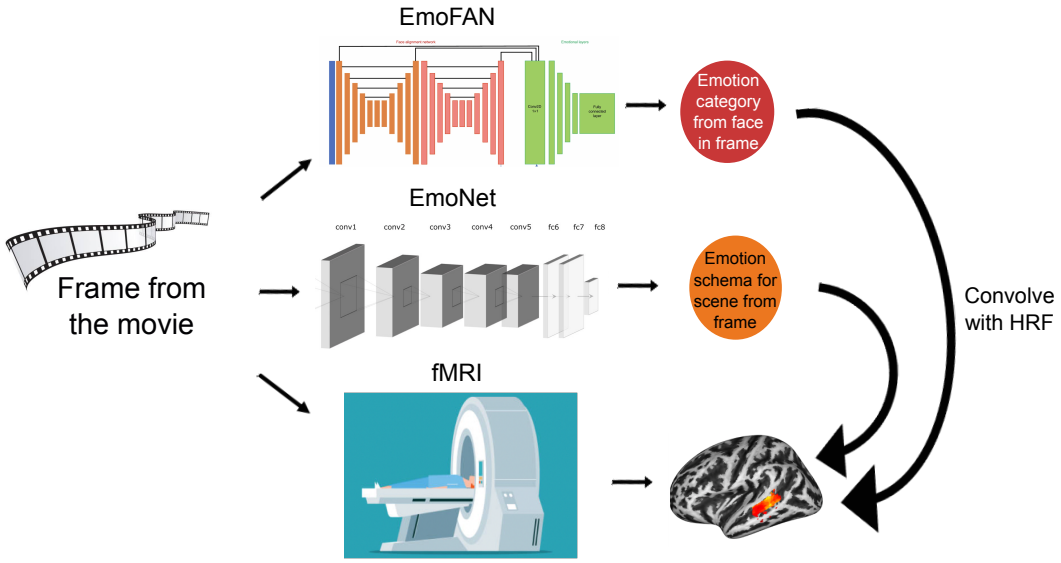

Figure 1. Encoding models for facial expressions and visual context. Every fifth frame of the movie is fed into two artificial neural networks. The upper ANN ("EmoFAN") identifies emotional facial expressions, while the lower ANN ("EmoNet") classifies the emotion schema of the image. Activations from layers of each of these ANNs are convolved with the hemodynamic response function (HRF) to create an encoding model of predicted brain activity based on facial expression features for EmoFAN and features related to visual context for EmoNet. These variables are used as features in multivariate encoding models to predict BOLD activation in the brains of subjects watching the full-length movie.

Given evidence of the distinct contributions of the amygdala and posterior STS to face processing, we hypothesized that posterior STS encodes representations uniquely related to facial expressions, and that the amygdala encodes representations related to the emotional significance of visual context more broadly. If this is the case, we would expect a double dissociation, such that abstract representations of facial expressions from EmoFAN would meaningfully predict activity in posterior STS and not the amygdala, with the opposite pattern for abstract emotion representations in EmoNet. However, given the evidence that the posterior STS responds to visual signals of emotion and may contain representations of emotion categories over and above those related to specific facial expressions, visual context (captured by EmoNet) might alternatively be encoded in posterior STS as well as the amygdala.

To evaluate these alternative accounts, we developed multivariate encoding models using features from the final layer of each ANN; for EmoFAN, this consisted of a 10-dimensional layer capturing the probabilities of 8 emotion categories as well as continuous variables for valence and arousal; for EmoNet, this consisted of a 20-dimensional layer of 20 emotion categories (for more detail on network architectures, see Methods). We first evaluated whether features related to facial expressions are encoded in posterior STS over and above variables related to visual context, also considering the possibility that emotional information from multiple sources are jointly encoded in posterior STS.

We found evidence that both encoding models—the one based on facial expressions and the one based on visual context—explained activity in posterior STS (average EmoFAN prediction-outcome correlation = .033, *SD* = .011, 95% CI = [.0282 to .0378], Cohen's *d* = 3.12; average EmoNet prediction-outcome correlation = .086, *SD* = .018, 95% CI = [.078 to .094],

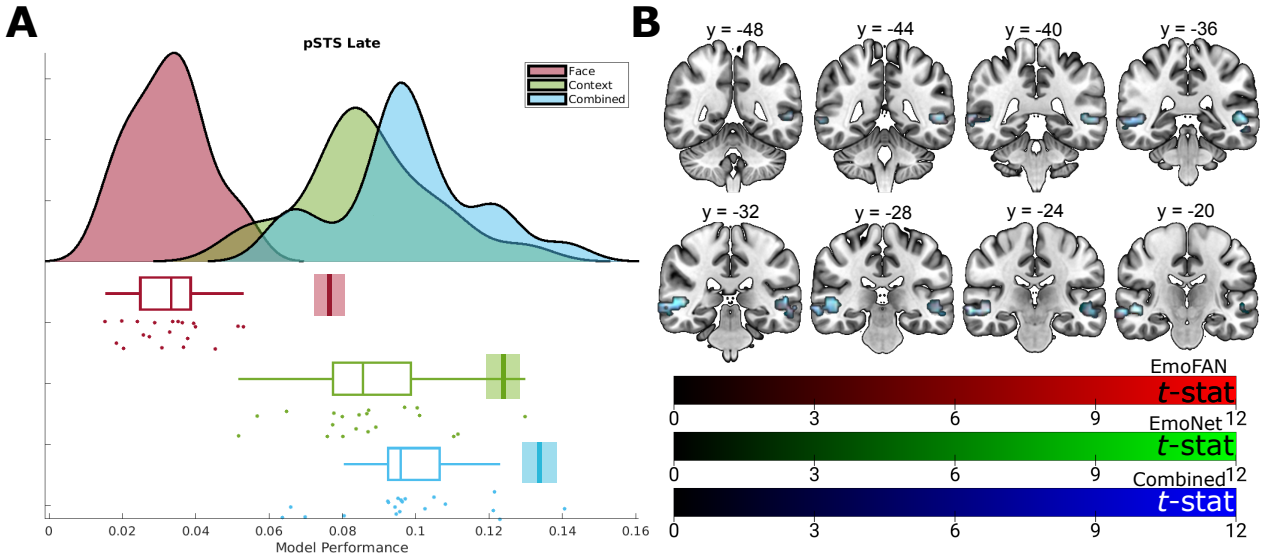

Figure 2. (**A**) Encoding model performance for features from the late layers of each model predicting posterior STS activity. EmoFAN shown in maroon, EmoNet in green, and the combined model in blue. Density and boxplots show model performance across subjects; each dot represents that model's performance on one subject's brain data. (**B**) Voxels significantly predicted by EmoFAN, EmoNet, and the combined model are overlapping throughout the posterior STS. Overlay plot shows EmoFAN vs. Chance in red, EmoNet vs. Chance in green, and the combined model vs. Chance in blue. Estimated noise ceilings determined by resubstitution are indicated by the vertical bars (standard error in lighter shade).

Cohen's d = 4.70; Figure 2). Voxelwise inference (one-sample t-test, FDR $q < .05$, two-tailed) revealed that a broad extent of voxels in bilateral posterior STS were significantly predicted by both EmoFAN (peak effect in dorsal left posterior STS: extent = 269 voxels/8808 mm^3, t = 9.71, location in MNI space = [-56, -35, -2]; peak effect in ventral right posterior STS: extent = 357 voxels/11688 mm^3, t = 12.64, location in MNI space = [56, -29, -5]) and EmoNet (peak effect in dorsal left posterior STS: extent = 283 voxels/9264 mm^3, t = 16.94, location in MNI space = [-56, -35, -2]; peak effect in ventral right posterior STS: extent = 361 voxels/11824 mm^3, t = 15.71, location in MNI space = [56, -29, -5]).

Because the amygdala has been implicated in rapid processing of certain facial features, particularly those related to rapid orienting towards threat-relevant emotions like anger and fear[26,43] rather than multiple abstract categories, we hypothesized that abstract representations in later layers of EmoFAN would not be directly encoded in patterns of amygdala activity. We tested this prediction by evaluating whether the encoding model using abstract representations from EmoNet would predict amygdala activity better than the model using abstract representations of facial expressions from EmoFAN. The results supported our hypothesis; the encoding model derived from EmoNet, but not the one derived from EmoFAN, explained response patterns in the amygdala (average EmoNet prediction-outcome correlation = .0094, SD = .0099, 95% CI = [-.034 to .0527], Cohen's d = .949; average EmoFAN prediction-outcome correlation < .001, SD = .0074, 95% CI = [-.0026 to .0038]; Figure 3a).

Voxelwise inference revealed that bilateral basolateral amygdalae were predicted by EmoNet (Figure 3b; peak effect in right amygdala: extent = 87 voxels/2848 mm^3, t = 7.65, location in MNI space = [23, -2, -17]; peak effect in left amygdala: extent = 70 voxels/2288 mm^3, t = 7.05, location in MNI space = [-20, -2, -17]). In addition, when comparing EmoNet and

EmoFAN directly, voxels in the basolateral amygdala and several in the superficial amygdala were significantly better predicted by EmoNet (Supplementary Figure 1; peak effect in right

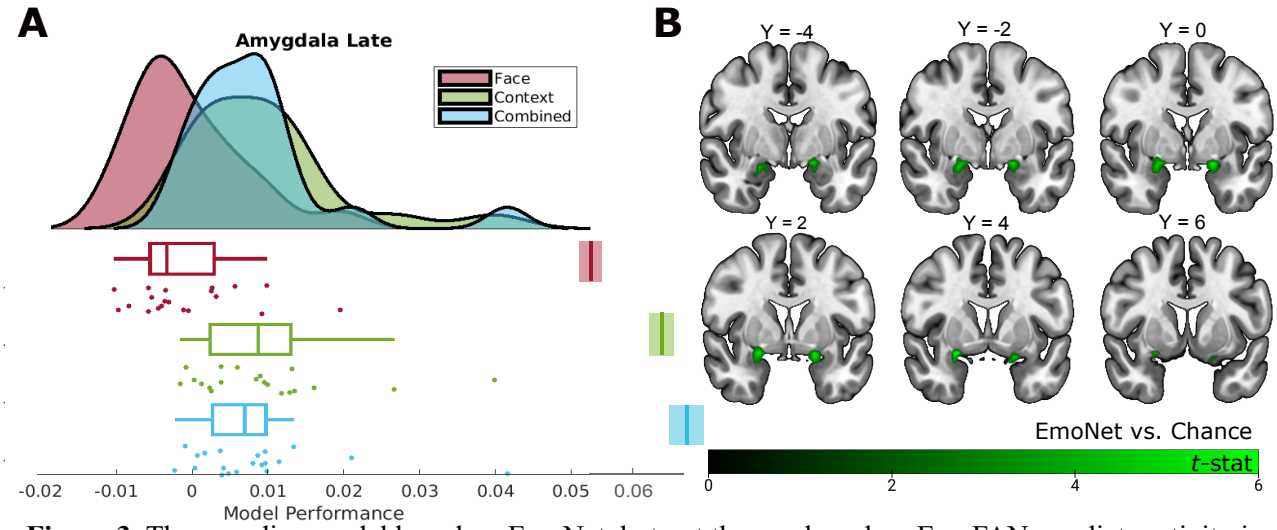

Figure 3. The encoding model based on EmoNet, but not the one based on EmoFAN, predicts activity in the amygdala. (**A**) Encoding model performance for features from the late layers of each model predicting amygdala activity. EmoFAN is shown in maroon, EmoNet is shown in green, and the combined model is shown in blue. Density and boxplots show model performance across subjects; each dot represents that model's performance on one subject's brain data. (**B**) Voxels for which activation was significantly explained by the EmoNet encoding model. Estimated noise ceilings determined by resubstitution are indicated by the vertical bars (standard error in lighter shade).

superficial amygdala: extent = 7 voxels/216 mm^3, $t = 4.22$, location in MNI space = [17, -5, -20]; peak effect in left superficial amygdala: extent = 6 voxels/208 mm^3, $t = 3.98$, location in MNI space = [-14, -5, -17]; peak effect in left basolateral amygdala: extent = 47 voxels/1536 mm^3, $t = 6.12$, location in MNI space = [-20, -2, -14]).

To determine whether representations of facial expressions and visual context each explain unique variance in posterior STS activity, we created a joint encoding model for each region that combines features from both ANNs. We did this by using features from both models as predictors of brain activity; if the information in the two models is redundant, the combined model should do no better than either individual model, whereas if each model contains distinct information, the combined model should have a higher predictive value. To examine whether this joint model performed differently across regions (i.e., that the combined model was significantly

more predictive of posterior STS than amygdala activity), we performed a 2-way ANOVA and found a significant region by model interaction ($F_{2,38} = 127.11$, $p < .001$, partial $\eta^2 = .870$). The joint encoding model explained significantly more posterior STS activation than either model alone (average prediction-outcome correlation = .098, SD = .019, 95% CI = [.089 to .1060], Cohen's d = 5.09), suggesting that each of these sources of information are uniquely encoded in posterior STS. This pattern of effects was not present for the amygdala, as the combined model did not explain more activity compared to the visual context model alone (average prediction-outcome correlation = .0082, SD = .0096, 95% CI = [.004 to .012], Cohen's d = .857).

To determine which voxels in posterior STS were better predicted by the combined model compared to either EmoFAN or EmoNet alone, we performed group inference across subjects and found that voxels in bilateral posterior STS were better predicted by the combined model compared to either model alone (Supplementary Figure 2; peak effect in left dorsal posterior STS: extent = 285 voxels/9328 mm^3, $t = 23.28$, location in MNI space = [-56, -35, -2]; peak effect in right ventral posterior STS: extent = 361 voxels/11824 mm^3, $t = 18.10$, location in MNI space = [56, -29, -5]).

Because the amygdala and posterior STS are known to respond to basic visual features of faces, including the high contrast present in the eyes and low frequency information about face shape[11,44,45], it is possible that activity in earlier layers of ANNs—which are ultimately transformed into representations of emotion categories—might predict responses in these areas. To examine the predictive ability of encoding models controlling for these more basic features, we fit a new set of encoding models based on each ANN's intermediate features, taken from the activations of layers earlier in the stream of processing. We found that both of these encoding models significantly explained activation in the amygdala (Figure 4; average EmoFAN

prediction-outcome correlation = .037, SD = .018, 95% CI = [.029 to .045], Cohen's *d* = 2.1; average EmoNet prediction-outcome correlation = .046, SD = .023, 95% CI = [.036 to .056], Cohen's *d* = 2.04) and posterior STS (average EmoFAN prediction-outcome correlation = .172, SD = .034, 95% CI = [.157 to .187], Cohen's *d* = 5.12; average EmoNet prediction-outcome correlation = .236, SD = .043, 95% CI = [.217 to .255], Cohen's *d* = 5.51). As intermediate layers of the ANNs represent earlier stages of visual processing, this demonstrates that more basic visual features related to emotional expressions and visual context explain activity in both regions.

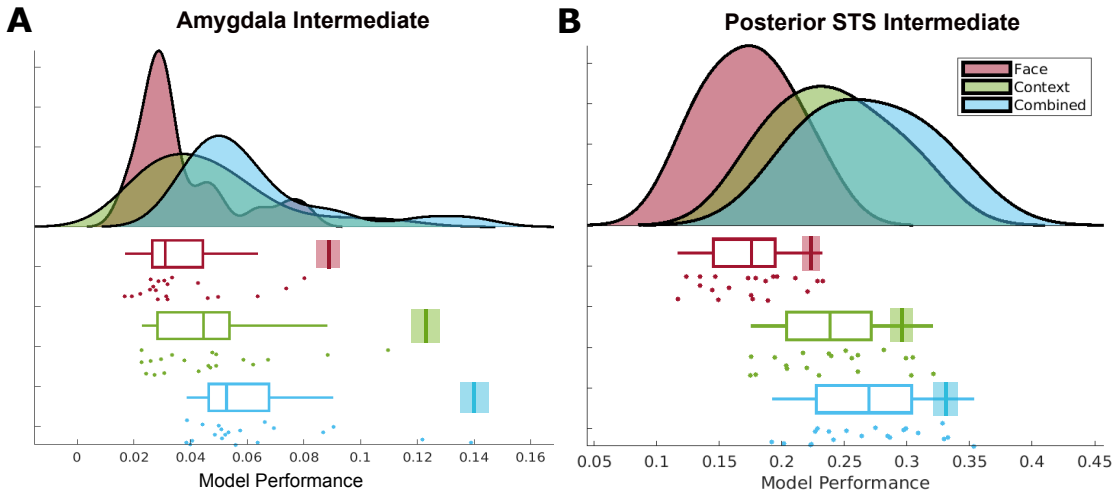

Figure 4. Intermediate layers of both encoding models, as well as the combined model, predict activity in the amygdala and posterior STS. Encoding model performance for features from the late layers of each model predicting activity in the region for (**A**) amygdala and (**B**) posterior STS. EmoFAN is shown in maroon, EmoNet is shown in green, and the combined model is shown in blue. Density and boxplots show model performance across subjects; each dot represents that model's performance on one subject's brain data. Estimated noise ceilings determined by resubstitution are indicated by the vertical bars (standard error in lighter shade).

To compare the effects of depth (intermediate vs. late layer), region (amygdala vs. posterior STS), and model (EmoFAN, EmoNet, and combined), we performed a 3-way ANOVA and found a significant region by depth by model interaction ($F_{2,38} = 7.42$, $p = .005$, partial $\eta^2 = .281$). This analysis revealed that although earlier representations from an intermediate layer of EmoFAN predict amygdala and posterior STS activity, as representations in EmoFAN become

more specialized for the particular objective of categorizing a facial expression, they no longer meaningfully explain amygdala responses ($r < .001$). To examine the effect of encoding model for models based on intermediate layers, we performed a follow-up 2-way ANOVA and found a significant region by model interaction for intermediate layers: ($F_{2,38} = 284.13$, $p < .001$, partial $\eta^2 = .937$). This interaction was driven by a larger effect of model type in posterior STS, such that the combined model showed a stronger additive effect (i.e., more variance explained features from EmoFAN) in posterior STS compared to the amygdala. This demonstrates that abstract representations of facial emotions are not good predictors of amygdala responses to naturalistic videos compared to more basic facial features or visual features that extend beyond the face.

Discussion

This work provides a more precise understanding of how visual information about facial expressions and emotional context is encoded in the brain. We found that posterior STS encodes abstract emotion representations related specifically to emotional facial expressions, while the amygdala does not. We also found that abstract emotion representations related to visual context are encoded in both amygdala and posterior STS. Furthermore, an encoding model combining features related to both facial expression and visual context explained a greater degree of posterior STS activation than either model alone. These findings suggest that the posterior STS processes visual information related to emotion from both facial expressions and the broader visual context, while the amygdala may represent visual context rather than specific categories derived from facial expressions. In addition, this work suggests that ANNs with varying objectives, including face localization, expression recognition, and schema classification, can explain activity in brain regions underlying social perception.

These results have implications for cognitive models of face perception: in particular, how the amygdala and posterior STS contribute to the distributed network of brain regions that processes faces. By isolating features from faces enmeshed in a variety of visual contexts, we observed that features derived from facial expressions did not predict amygdala activity whereas features from the broader visual context did. Furthermore, features from intermediate but not late layers of EmoFAN predicted amygdala activity, demonstrating that the amygdala responds to more basic visual representations that could be used to classify facial expressions. This is consistent with evidence suggesting that the amygdala receives information about low spatial frequency components of faces to orient to the eye region and guide behavioral responses to threat and reward[29,44–46]. The final layer of EmoFAN did not predict amygdala activity; in the process of calculating what emotion category a face belongs to, EmoFAN abstracts away from such low-level features and information that predicts amygdala activity is lost. From this view, the amygdala responds to emotional faces because they are salient indicators of important events, not because it represents the emotion category of a facial expression in an abstract way.

Additionally, our results shed light on the role of posterior STS in face processing. We found that abstract emotion features derived from both facial expressions and visual context predict unique components of posterior STS activity; that is, the joint encoding model based on late layers from both EmoFAN and EmoNet predicted a greater proportion of posterior STS activity than either model alone. If posterior STS were only encoding abstract emotion categories, we would expect the combined model to perform no better than the single models. We thus conclude that it is unlikely that the only function of posterior STS is to represent abstract emotion categories that have been completely disentangled from their sensory source.

We further found that intermediate and late layers of both models—facial expression and emotional context—uniquely predicted posterior STS activity during naturalistic movie-viewing. This improvement in prediction suggests that posterior STS encodes both types of representations: those that are more specific to the visual modality, as well as those that are more abstract and tied to emotion categories. Debate continues over which set of variables are encoded in this region—whether it is specialized for processing facial movement or whether it takes modality-specific information as input and maps it to supramodal representations of emotion categories[16,47]. The results of our study support the idea that neural populations in posterior STS encode both types of representations: those associated with specific modalities and supramodal abstract emotion categorizations.

This study demonstrates the utility of using ANNs as models of perceptual systems beyond object recognition in the ventral visual stream. Examining the mappings between ANNs and brain activity allows for a finer grained understanding of the neural computations involved in transforming visual input into representations that could be used to recognize emotion in the face. The ANNs we examined here have different objectives and architectures, with EmoFAN standing apart from most deep convolutional networks including those that categorize faces[48] by including an initial filter that detects the current pose of a face and identifies important landmarks, which then influence a set of convolutional layers that operate on facial features at progressively lower dimensionalities. With this relatively straightforward architecture, we see that an EmoFAN-derived encoding model significantly explains posterior STS activity. Although the amount of variance explained is far from explaining all activity in this region, and the structure of EmoFAN is far simpler than that of the face processing systems in humans, the representations that EmoFAN has learned are meaningfully related to human brain activity, and

thus advance our understanding of social information processing in the brain. To obtain a model that better approximates posterior STS function, alterations can be made that account for other operations occurring in this region, including processing auditory and linguistic features. To understand how posterior STS processes social and emotional information more fully, future work should explore how features from other modalities might improve the explanatory power of an encoding model predicting posterior STS activity.

This work is among a number of studies employing naturalistic paradigms to study social perception[49–51]. Naturalistic approaches allow us to examine how the brain represents the multifaceted, complex nature of emotion; however, it is challenging to analyze experiments of this type. Stimuli are not presented in a controlled manner, which makes it difficult to isolate the effects of any one factor or component. Our approach, in which ANNs are used to extract complex features related to emotion, allows multifaceted stimuli to be parsed and corresponding neural responses to be probed. In this framework, any neural network can be used to extract features related to the task it has been trained for, and these features can be tested for the degree to which they map on to brain activity[52]. Importantly, this allows us to better understand the brain as it is functioning in a more ecologically valid context. In the real world, it is rare that humans perceive others' emotions in a vacuum, and we can better capture such contextually embedded processes using the present approach.

This work highlights multiple fruitful directions for future research. Although EmoFAN and EmoNet are able to accomplish the tasks they were trained for, they each constitute only one way of solving the given problem that stems from how they were trained; there exists a wide space of neural networks that accomplish similar tasks. EmoFAN, for example, was trained to classify the category of static images of annotated facial expressions and is applied here on

frames from a full-length movie. The facial expressions in the movie are dynamic, so it is likely that a neural network trained to learn dynamic sequences of facial expressions[53] would better capture relevant features. In addition, subjects watching the movie were not tasked with identifying the emotions from the faces they saw, and while emotional faces are likely attended to due to their salience[54], we did not evaluate brain responses to face processing during an explicit recognition task. Additionally, the sample consists of 20 subjects, which is insufficient to characterize any sex-related, cultural, developmental, or racial impacts on how the brain processes emotional faces. In addition, this study does not causally test how information is conveyed between the amygdala and posterior STS. Some work has found evidence that they are causally connected by showing that transcranial magnetic stimulation applied over STS during face perception leads to a reduced amygdala response[55]. Here we do not test whether information is transmitted between these regions.

The overall proportion of brain activity explained by encoding models (i.e., the noise ceiling) in the pSTS and amygdala was small, likely as a consequence of the constrained nature of each ANN and the complexity of the naturalistic free-viewing paradigm. This illustrates how detecting emotion conveyed by visual signals is but one of many functions performed by these regions. Past work has used tightly controlled experimental paradigms that aim to minimize context effects and variation in brain responses over time[32,56]. Experiments of this kind minimize noise, but they may miss out on meaningful aspects of brain function, including multimodal processing and integration that occur in more ecologically valid paradigms. It is likely that more complex models, trained on a range of relevant tasks with inputs from multiple modalities, will be necessary to fully explain the function of these brain regions during emotion perception. Future modeling efforts could focus on the relationships between these and other regions to

further illuminate the distributed brain network involved in processing emotional visual information.

By testing ANN-derived encoding models on brain activity during naturalistic movie viewing, this study sheds light on how signals of emotion are represented in the brain. The emotional meaning of the environment can be sensed from multiple sources, including others' facial expressions and visual cues related to emotional context. Our findings suggest that although the amygdala may represent salience and emotional context, it does not strongly encode specific facial expressions; the posterior STS operates on both facial expression-specific representations and abstractions from sensory inputs to emotion categories. Further exploration using neural network-derived encoding models will deepen our knowledge—not only *that* emotion signals are processed in these regions, but *how* they convert visual inputs into an abstract representation of the present emotion state. Doing so will increase our understanding of the neural computations that enable the human brain to take in the complex emotional world.

Methods

fMRI Data

Data from the Naturalistic Neuroimaging Database[42] was used for this study. In this database, participants watched a range of full-length movies that contained rich social and emotional content. Their only task was to watch the movie, which makes this data ideal for investigating how signals of emotion might be encoded in naturalistic contexts.

Participants

20 subjects were recruited from the London area (mean age = 27.7, 50% female, 30% Black, Asian, or ethnic minority). Participants were screened out if they had previously seen the

movie (*500 Days of Summer*), so that all participants were viewing it for the first time when they completed the study. In addition, all participants were right-handed and native English speakers. Participants were excluded if they had a history of claustrophobia, psychiatric or neurological illness, if they were taking medication, or if they had a hearing or uncorrected visual impairment.

Paradigm

Brain activity was measured using fMRI while subjects viewed the full-length movie *500 Days of Summer*. The movie was presented in two ~50 minute segments. Participants were able to pause in the middle of a segment if they needed a break, after which the scan and the movie were resumed in a synchronized manner (See "Movie pausing" in Aliko et al., 2020). After the movie was complete, an anatomical scan was acquired.

MRI acquisition

Data was acquired on a 1.5 T Siemens MAGNETOM Avanto with a 32 channel head coil (Siemens Healthcare, Erlangen, Germany). For functional images during movie-viewing, a multiband EPI sequence was used (TR = 1 s, TE = 54.8 ms, flip angle of 75°, 40 interleaved slices, resolution = 3.2 mm isotropic). For the anatomical image, a 10-minute high-resolution T1-weighted MPRAGE was collected (R = 2.73 s, TE = 3.57 ms, 176 sagittal slices, resolution = 1.0 mm).

Preprocessing

Preprocessing was carried out through AFNI and included slice-time corrections, de-spiking, volume registration, alignment to the MNI template, 6 mm smoothing, detrending with 6 motion regressors, timing correction to account for pauses in scanning, concatenation of runs

across the movie, and manual artifact removal with ICA (for a detailed description of all preprocessing steps, see "Preprocessing: Functional" in Aliko et al., 2020).

Definition of Regions of Interest

Regions of interest representing the posterior STS and amygdala were obtained from cortical and subcortical atlases, respectively. For posterior STS, the dorsal and ventral portions of posterior STS ("STSdp" and "STSvp") from a multimodal parcellation of the human cortex were selected[57]. For the amygdala, anatomical subdivisions based on cytoarchitecture[58] were selected from the SPM anatomy toolbox (including superficial, basolateral, and centromedial).

Analyses

Encoding Model Creation from ANNs

To test whether representations of facial expression category or visual emotion schema are encoded in brain activity in the amygdala and posterior STS, we used two ANNs to process the movie and generate encoding models. One, called EmoFAN, is a deep convolutional neural network trained to classify the emotional expression of faces[39]. EmoFAN uses a face-alignment network[59] to identify facial landmarks, which are then applied through an attention mechanism to a set of four convolutional blocks (each containing three convolutional layers). The final layer of EmoFAN has 10 dimensions, and contains probabilities that the current stimulus falls into 1 of 8 emotion categories (neutral, happy, sad, surprise, fear, disgust, anger, and contempt) as well as continuous values for valence and arousal. The other, named EmoNet, is a deep convolutional neural network trained to classify the emotion schema evoked by a frame of visual input—i.e., visual context. EmoNet is a convolutional neural network based on AlexNet that consists of five convolutional layers and three fully connected layers, and was trained to identify the human-

rated emotional state of images from emotional video clips (for emotion categories, see Kragel et al., 2019). To develop the encoding models, we fed every fifth frame of the movie into each ANN and extracted activations from intermediate and late layers of each model. The intermediate layer for EmoFAN was the concatenated convolutional layers of the final convolutional block, while the intermediate layer for EmoNet was the penultimate fully connected layer; the late layers were the last fully connected layer for both ANNs.

Comparison with Brain Data

Having obtained features from both ANNs for each frame of the video, we next convolved these features with the hemodynamic response function to generate an expected response if a given brain region was encoding these features. Using these features, we used partial least squares (PLS) regression to search for a model that explained BOLD activity in the region of interest. For EmoFAN, a 10-dimensional PLS regression model was used; for EmoNet, 20 dimensions were used; for the combined model, 20 dimensions were used. Using 5-fold cross validation, we trained the model on four-fifths of the subjects' data and tested on the remaining fifth by calculating the correlation between predicted and actual brain activity.

Noise Ceiling Calculations

To estimate the noise ceiling for each region of interest, we performed PLS regression using resubstitution. That is, we trained a PLS regression model on all of each subject's data and tested it on the same data, which provides an upper bound for how much variance a model of a given complexity would ever be able to explain. For posterior STS response, the resubstituted prediction-outcome correlation averaged across 20 subjects is .0784 (SE=.002) for the late layer of EmoFAN, .121 (SE=.003) for the late layer of EmoNet, and .134 (SE=.003) for the late-layer

combined model; it is .213 (SE=.007) for the intermediate layer of EmoFAN, .293 (SE=.009) for the intermediate layer of EmoNet, and .319 (SE=.010) for the intermediate-layer combined model. For amygdala response, the resubstituted prediction-outcome correlation averaged across 20 subjects is .048 (SE=.001) for the late layer of EmoFAN, .066 (SE=.001) for the late layer of EmoNet, .072 (SE=.002) for the late-layer combined model; it is .087 (SE=.004) for the intermediate layer of EmoFAN, .123 (SE=.005) for the intermediate layer of EmoNet, and .140 (SE=.005) for the intermediate-layer combined model. Noise ceiling estimates are indicated in the figures with a vertical line.

Statistical Tests

First, we performed t-tests to determine whether the average prediction-outcome correlation was significantly above zero. Next, we analyzed the model weights (B_{PLS}) to determine which voxels were significantly predicted by our encoding models across subjects. A false discovery rate correction of $q = .05$ was used as a threshold to determine significance. To compare the effects of model, region, and depth, we performed ANOVAs on the average prediction-outcome correlation across subjects. We performed two-way ANOVAs to examine the effect of model and region, and we did this for late and intermediate layers. In addition, we performed a three-way ANOVA to test the effects of model, region, and depth all together.

References

1. Bruce, V. & Young, A. Understanding face recognition. *Br. J. Psychol.* **77**, 305–327 (1986).
2. Haxby, J. V., Hoffman, E. A. & Gobbini, M. I. The distributed human neural system for face perception. *Trends Cogn. Sci.* **4**, 223–233 (2000).
3. Calder, A. J. & Young, A. W. Understanding the recognition of facial identity and facial expression. *Nat. Rev. Neurosci.* **6**, 641–651 (2005).
4. Duchaine, B. & Yovel, G. A revised neural framework for face processing. *Annu. Rev. Vis. Sci.* **1**, 393–416 (2015).
5. Said, C. P., Moore, C. D., Norman, K., Haxby, J. V. & Todorov, A. Graded representations of emotional expressions in the left superior temporal sulcus. *Front. Syst. Neurosci.* **4**, (2010).
6. Deen, B., Koldewyn, K., Kanwisher, N. & Saxe, R. Functional organization of social perception and cognition in the superior temporal sulcus. *Cereb. Cortex* **25**, 4596–4609 (2015).
7. Wegrzyn, M. *et al.* Investigating the brain basis of facial expression perception using multi-voxel pattern analysis. *Cortex* **69**, 131–140 (2015).
8. LaBar, K. S., Crupain, M. J., Voyvodic, J. T. & McCarthy, G. Dynamic perception of facial affect and identity in the human brain. *Cereb. Cortex* **13**, 1023–1033 (2003).
9. Johnson, M. H. Subcortical face processing. *Nat. Rev. Neurosci.* **6**, 766–774 (2005).
10. Zhang, H. *et al.* Face-selective regions differ in their ability to classify facial expressions. *NeuroImage* **130**, 77–90 (2016).
11. Grossman, E. D. & Blake, R. Brain areas active during visual perception of biological motion. *Neuron* **35**, 1167–1175 (2002).

12. Pelphrey, K. A., Morris, J. P., Michelich, C. R., Allison, T. & McCarthy, G. Functional anatomy of biological motion perception in posterior temporal cortex: An fMRI study of eye, mouth and hand Movements. *Cereb. Cortex* **15**, 1866–1876 (2005).
13. Said, C. P., Moore, C. D., Engell, A. D., Todorov, A. & Haxby, J. V. Distributed representations of dynamic facial expressions in the superior temporal sulcus. *J. Vis.* **10**, 11–11 (2010).
14. Peelen, M. V., Atkinson, A. P. & Vuilleumier, P. Supramodal representations of perceived emotions in the human brain. *J. Neurosci.* **30**, 10127–10134 (2010).
15. Watson, R. *et al.* Crossmodal adaptation in right posterior superior temporal sulcus during face–voice emotional integration. *J. Neurosci.* **34**, 6813–6821 (2014).
16. Schirmer, A. & Adolphs, R. Emotion perception from face, voice, and touch: Comparisons and convergence. *Trends Cogn. Sci.* **21**, 216–228 (2017).
17. Wilensky, A. E., Schafe, G. E., Kristensen, M. P. & LeDoux, J. E. Rethinking the fear circuit: The central nucleus of the amygdala Is required for the acquisition, consolidation, and expression of Pavlovian fear conditioning. *J. Neurosci.* **26**, 12387–12396 (2006).
18. Öhman, A., Carlsson, K., Lundqvist, D. & Ingvar, M. On the unconscious subcortical origin of human fear. *Physiol. Behav.* **92**, 180–185 (2007).
19. Hooker, C. I., Germine, L. T., Knight, R. T. & D'Esposito, M. Amygdala response to facial expressions reflects emotional learning. *J. Neurosci.* **26**, 8915–8922 (2006).
20. Graham, R. & LaBar, K. S. Neurocognitive mechanisms of gaze-expression interactions in face processing and social attention. *Neuropsychologia* **50**, 553–566 (2012).

21. Fitzgerald, D. A., Angstadt, M., Jelsone, L. M., Nathan, P. J. & Phan, K. L. Beyond threat: Amygdala reactivity across multiple expressions of facial affect. *NeuroImage* **30**, 1441–1448 (2006).
22. Phillips, M. L. *et al.* Neural responses to facial and vocal expressions of fear and disgust. *Proc. R. Soc. Lond. B Biol. Sci.* **265**, 1809–1817 (1998).
23. Loughead, J., Gur, R. C., Elliott, M. & Gur, R. E. Neural circuitry for accurate identification of facial emotions. *Brain Res.* **1194**, 37–44 (2008).
24. Mattavelli, G. *et al.* Neural responses to facial expressions support the role of the amygdala in processing threat. *Soc. Cogn. Affect. Neurosci.* **9**, 1684–1689 (2014).
25. Van Der Gaag, C., Minderaa, R. B. & Keysers, C. The BOLD signal in the amygdala does not differentiate between dynamic facial expressions. *Soc. Cogn. Affect. Neurosci.* **2**, 93–103 (2007).
26. Wang, S. *et al.* The human amygdala parametrically encodes the intensity of specific facial emotions and their categorical ambiguity. *Nat. Commun.* **8**, 14821 (2017).
27. Öhman, A. The role of the amygdala in human fear: Automatic detection of threat. *Psychoneuroendocrinology* **30**, 953–958 (2005).
28. Sander, D., Grafman, J. & Zalla, T. The human amygdala: An evolved system for relevance detection. *Rev. Neurosci.* **14**, (2003).
29. Adolphs, R. Fear, faces, and the human amygdala. *Curr. Opin. Neurobiol.* **18**, 166–172 (2008).
30. Hariri, A. R., Tessitore, A., Mattay, V. S., Fera, F. & Weinberger, D. R. The amygdala response to emotional stimuli: A comparison of faces and scenes. *NeuroImage* **17**, 317–323 (2002).

31. Yamins, D. L. K. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci.* **111**, 8619–8624 (2014).
32. Khaligh-Razavi, S.-M. & Kriegeskorte, N. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* **10**, e1003915 (2014).
33. Dobs, K., Isik, L., Pantazis, D. & Kanwisher, N. How face perception unfolds over time. *Nat. Commun.* **10**, 1258 (2019).
34. Baek, S., Song, M., Jang, J., Kim, G. & Paik, S.-B. Face detection in untrained deep neural networks. *Nat. Commun.* **12**, 7328 (2021).
35. Kanwisher, N., Khosla, M. & Dobs, K. Using artificial neural networks to ask 'why' questions of minds and brains. *Trends Neurosci.* **46**, 240–254 (2023).
36. Dobs, K., Martinez, J., Kell, A. J. E. & Kanwisher, N. Brain-like functional specialization emerges spontaneously in deep neural networks. *Sci. Adv.* **8**, eabl8913 (2022).
37. Pitcher, D. & Ungerleider, L. G. Evidence for a third visual pathway specialized for social perception. *Trends Cogn. Sci.* **25**, 100–110 (2021).
38. Sonkusare, S., Breakspear, M. & Guo, C. Naturalistic stimuli in neuroscience: Critically acclaimed. *Trends Cogn. Sci.* **23**, 699–714 (2019).
39. Toisoul, A., Kossaifi, J., Bulat, A., Tzimiropoulos, G. & Pantic, M. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nat. Mach. Intell.* **3**, 42–50 (2021).
40. Kragel, P. A., Reddan, M. C., LaBar, K. S. & Wager, T. D. Emotion schemas are embedded in the human visual system. *Sci. Adv.* **5**, eaaw4358 (2019).
41. Naselaris, T., Kay, K. N., Nishimoto, S. & Gallant, J. L. Encoding and decoding in fMRI. *NeuroImage* **56**, 400–410 (2011).

42. Aliko, S., Huang, J., Gheorghiu, F., Meliss, S. & Skipper, J. I. A naturalistic neuroimaging database for understanding the brain using ecological stimuli. *Sci. Data* **7**, 347 (2020).
43. Adolphs, R. *et al.* A mechanism for impaired fear recognition after amygdala damage. *Nature* **433**, 68–72 (2005).
44. Whalen, P. J. *et al.* Human amygdala responsivity to masked fearful eye whites. *Science* **306**, 2061–2061 (2004).
45. Méndez-Bértolo, C. *et al.* A fast pathway for fear in human amygdala. *Nat. Neurosci.* **19**, 1041–1049 (2016).
46. Gosselin, F., Spezio, M. L., Tranel, D. & Adolphs, R. Asymmetrical use of eye information from faces following unilateral amygdala damage. *Soc. Cogn. Affect. Neurosci.* **6**, 330–337 (2011).
47. Vaessen, M., Van der Heijden, K. & De Gelder, B. Decoding of emotion expression in the face, body and voice reveals sensory modality specific representations. *bioRxiv* (2019).
48. Parkhi, O., Vedaldi, A. & Zisserman, A. Deep face recognition. *Proc. Br. Mach. Vis. Conf.* (2015).
49. Jospe, K. *et al.* The contribution of linguistic and visual cues to physiological synchrony and empathic accuracy. *Cortex* **132**, 296–308 (2020).
50. Lee Masson, H. & Isik, L. Functional selectivity for social interaction perception in the human superior temporal sulcus during natural viewing. *NeuroImage* **245**, 118741 (2021).
51. Saarimäki, H. Naturalistic stimuli in affective neuroimaging: A review. *Front. Hum. Neurosci.* **15**, 675068 (2021).
52. Sievers, B. & Thornton, M. A. Deep social neuroscience: The promise and peril of using artificial neural networks to study the social brain. *PsyArXiv* (2023).

53. Snoek, L., Jack, R. E. & Schyns, P. E. Dynamic face imaging: A novel analysis framework for 4D social face perception and expression. *IEEE 17th Int. Conf. Autom. Face Gesture Recognit.* 1–4 (2023).
54. Rösler, L., End, A. & Gamer, M. Orienting towards social features in naturalistic scenes is reflexive. *PLOS ONE* **12**, e0182037 (2017).
55. Pitcher, D., Japee, S., Rauth, L. & Ungerleider, L. G. The superior temporal sulcus Is causally connected to the amygdala: A combined TBS-fMRI study. *J. Neurosci.* **37**, 1156–1161 (2017).
56. Allen, E. J. *et al.* A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nat. Neurosci.* **25**, 116–126 (2022).
57. Glasser, M. F. *et al.* A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171–178 (2016).
58. Amunts, K. *et al.* Cytoarchitectonic mapping of the human amygdala, hippocampal region and entorhinal cortex: intersubject variability and probability maps. *Anat. Embryol. (Berl.)* **210**, 343–352 (2005).
59. Bulat, A. & Tzimiropoulos, G. How far are we from solving the 2D & 3D face alignment problem? (And a dataset of 230,000 3D facial landmarks). *Proc. IEEE Int. Conf. Comput. Vis.* 1021–1030 (2017).

Author Contributions:

P.K. conceived of the study design; P.K., K.S., and G.J. analyzed and interpreted the data; K.S. wrote the manuscript; P.K. edited the manuscript.

Competing Interests statement:

The authors declare no competing interests.

Code Availability:

The code for this study will be available at the following GitHub repository:

https://github.com/ecco-laboratory/SEE