Article

Current Biology

Hierarchical organization of social action features along the lateral visual pathway

Highlights

- Naturalistic social actions reliably drive responses in the lateral visual pathway
- Features of social actions are hierarchically organized along the lateral pathway
- Communicative actions best predict activity in the superior temporal sulcus

Authors

Emalie McMahon, Michael F. Bonner, Leyla Isik

Correspondence

emaliemcmahon@jhu.edu

In brief

McMahon et al. find that features of naturalistic social actions become increasingly abstract along the posterior to anterior axis in the lateral visual pathway. They find that communicative actions uniquely explain responses in anterior regions of the lateral pathway, presenting evidence for a computational goal of the pathway.

McMahon et al., 2023, Current Biology *33*, 1–13
December 4, 2023
https://doi.org/10.1016/j.cub.2023.10.015

CellPress

Please cite this article in press as: McMahon et al., Hierarchical organization of social action features along the lateral visual pathway, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.10.015

Current Biology

CellPress
OPEN ACCESS

Article

Hierarchical organization of social action features along the lateral visual pathway

Emalie McMahon,[1,3,*] Michael F. Bonner,[1] and Leyla Isik[1,2]
[1]Department of Cognitive Science, Zanvyl Krieger School of Arts & Sciences, Johns Hopkins University, 237 Krieger Hall, 3400 N. Charles Street, Baltimore, MD 21218, USA
[2]Department of Biomedical Engineering, Whiting School of Engineering, Johns Hopkins University, Suite 400 West, Wyman Park Building, 3400 N. Charles Street, Baltimore, MD 21218, USA
[3]Lead contact
*Correspondence: emaliemcmahon@jhu.edu
https://doi.org/10.1016/j.cub.2023.10.015

SUMMARY

Recent theoretical work has argued that in addition to the classical ventral (what) and dorsal (where/how) visual streams, there is a third visual stream on the lateral surface of the brain specialized for processing social information. Like visual representations in the ventral and dorsal streams, representations in the lateral stream are thought to be hierarchically organized. However, no prior studies have comprehensively investigated the organization of naturalistic, social visual content in the lateral stream. To address this question, we curated a naturalistic stimulus set of 250 3-s videos of two people engaged in everyday actions. Each clip was richly annotated for its low-level visual features, mid-level scene and object properties, visual social primitives (including the distance between people and the extent to which they were facing), and high-level information about social interactions and affective content. Using a condition-rich fMRI experiment and a within-subject encoding model approach, we found that low-level visual features are represented in early visual cortex (EVC) and middle temporal (MT) area, mid-level visual social features in extrastriate body area (EBA) and lateral occipital complex (LOC), and high-level social interaction information along the superior temporal sulcus (STS). Communicative interactions, in particular, explained unique variance in regions of the STS after accounting for variance explained by all other labeled features. Taken together, these results provide support for representation of increasingly abstract social visual content—consistent with hierarchical organization—along the lateral visual stream and suggest that recognizing communicative actions may be a key computational goal of the lateral visual pathway.

INTRODUCTION

The ability to recognize people performing all kinds of activities is extremely important in our daily lives. One of the most common and important types of actions we see are social actions between two or more people,[1–3] like talking, hugging, or waving goodbye. Prior research has shown that the sociality of actions (i.e., the extent to which an action is directed at another person) is an important organizing feature of actions in the human brain,[1–3] particularly in the lateral occipital cortex. Recent work has also identified selective neural responses for dyadic social interactions in nearby regions along the superior temporal sulcus (STS) and homologous regions in the nonhuman primate brain.[4–6] This and related work have led to recent theoretical proposals for a third visual pathway on the lateral surface (in addition to the classic ventral and dorsal streams). The lateral visual stream is thought to be specialized for recognizing agentic action[7] or dynamic social perception more generally.[8]

As is characteristic of hierarchical organization of the ventral visual stream,[9] Pitcher and Ungerleider[8] argue that lateral pathway representations are organized from low-level features in the early visual cortex (EVC) to high-level features in the STS. However, this hierarchical organization and the specific features represented have not been comprehensively tested. In the current study, we aim to understand how social actions are organized in the brain and test the hypothesis that their features are extracted hierarchically along the lateral visual stream.

Recent work with simple stimuli has provided some evidence of increasingly abstract social action representations along the lateral surface. For instance, mid-level visual cues indicative of social actions, such as whether two bodies are facing[10,11] or moving toward one another,[12] are represented in the body-selective extrastriate body area (EBA) on the lateral surface. More anterior regions along the STS are selective for higher-level social action information, including the presence and valence of social interactions.[4,6,13–15]

Although these controlled studies have yielded important insights into neural selectivity for social content, naturalistic stimuli are critical for understanding human social perception due to the dynamic nature and extended temporal contingencies of social scenes.[16,17] Further, in order to understand the organization of a large region of the visual cortex that responds to many different visual and social features, it is important to broadly sample variance along many different dimensions, which is not possible to do comprehensively using controlled stimuli. For this reason, we opted for a condition-rich, naturalistic design.[3]

Please cite this article in press as: McMahon et al., Hierarchical organization of social action features along the lateral visual pathway, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.10.015

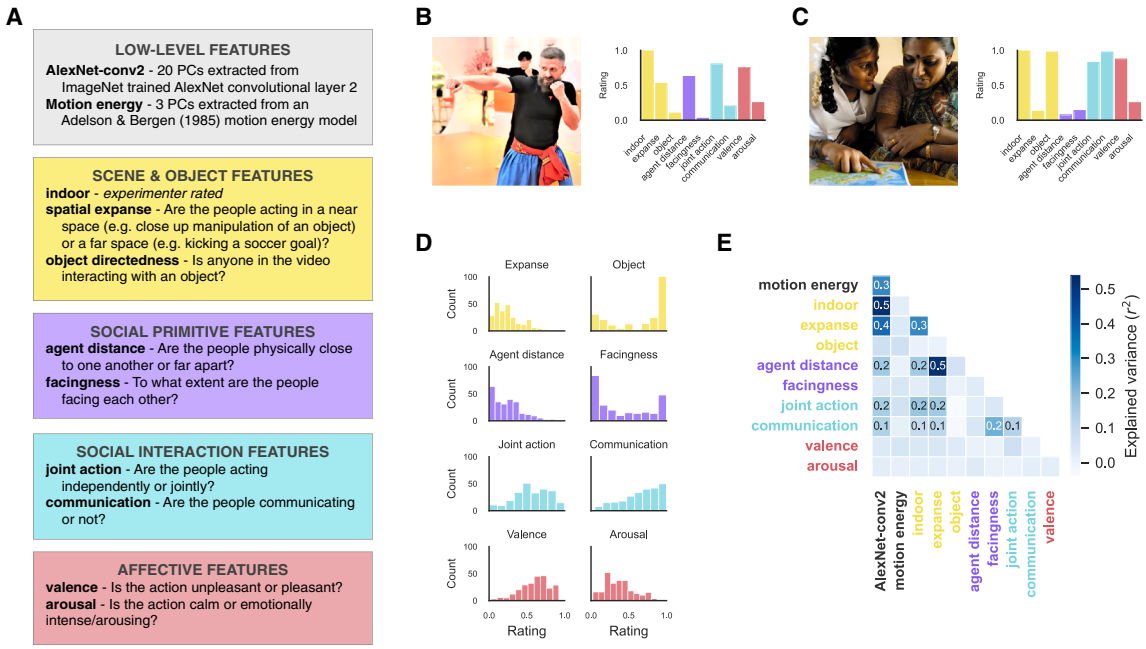

Figure 1. A richly annotated, naturalistic video dataset of dyadic social actions

(A) Labeled feature categories in our video dataset include low-level visual, scene and object, social primitive, social interaction, and affective features. The text description includes how the features were extracted algorithmically or the instructions that were presented to annotators.

(B and C) Images representative of videos in the dataset (one rated low and one rated high along the communication dimension), and the ratings on each of the annotated dimensions for the two videos.

(D) The distribution of ratings across all videos in the dataset for the annotated dimensions, except indoor because it is a binary dimension.

(E) The regression of each pair of features, where values in the cells indicate significant correlations determined by permutation testing (FDR $q < 0.05$).

Because of MiT license restrictions, images are only representative of videos in the set. (B) is "File:Oleksandr Chemykos Hopak Class 16 (49845104268).jpg" by Oleksandr Chemykos and (C) is "Parents and kids learn together" by DFID-UK Department for International Development. Both are licensed under CC BY 2.0.

We curated a novel large-scale, naturalistic stimulus set of dyadic social actions. These videos are dynamic and representative of real-world scenarios but still somewhat controlled: all videos contain exactly two people, and they were labeled and selected to ensure broad sampling of relevant visual and social features. We showed these videos to participants in a condition-rich fMRI experiment, which yielded high-quality data, particularly along the lateral surface of the brain. We used a voxel-wise encoding model approach to investigate how social action features were represented across the whole brain and in regions of interest (ROIs) along the lateral visual stream. We found that variance in posterior to anterior regions along the lateral surface was explained by increasingly complex social features. Specifically, communicative interactions uniquely explained variance along the STS above all other features, suggesting that recognizing communication (even in the absence of speech and language) is a key computational goal of the STS and lateral pathway.

RESULTS

A richly annotated, naturalistic video dataset of dyadic social actions

We curated a dataset of 250 two-person videos without sound from the Moments in Time action recognition dataset,[18] based on the video's action category and quality. Videos were each 3 s long and depicted typical, everyday actions based on responses to the American Time Use Survey.[19] We limited our stimulus set to two-person actions because the number of people in a scene is often correlated with low-level visual and higher-level social features.[1]

Each video was labeled on prior hypothesized features of social action and scene understanding. These feature categories included low-level visual features computed algorithmically (activations from AlexNet-conv2[20] and the output of a motion energy model[21]) and four categories of human-annotated features: scene and object features, visual social primitives, social interactions, and affective features (Figure 1A).

The first category of feature annotations included scene and object features. To understand a social event, it is important to establish the scene context in which it occurs and the objects that are involved, and recent proposals have suggested that these features may be critical to action representations along the lateral pathway.[7] Human annotators labeled three features in this category previously found to be important to action representations in the brain[2,3,22]: indoor scenes, the spatial expanse (or the spatial scale of the scene, such as a small bathroom versus a large auditorium), and object directedness (the extent to which people in the scene were interacting with objects).

High-level social information often correlates with visual features of people in a scene. In particular, the extent to which people are facing and the distance between them have been identified as two key features predicting whether people are judged as interacting.[23,24] However, two people can be nearby and facing one another but not be interacting, such as on crowded public

Please cite this article in press as: McMahon et al., Hierarchical organization of social action features along the lateral visual pathway, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.10.015

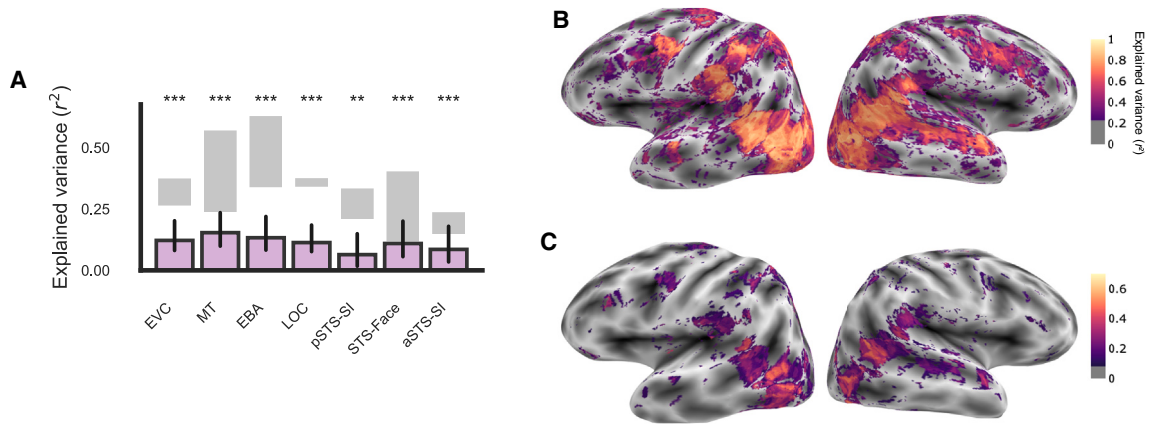

Figure 2. Extensive scanning yields high-quality fMRI data

(A) Group-level explained variance by the encoding model of all features (purple bars), calculated as the signed-squared correlation between the predicted and true responses in the test set. Error bars are the bootstrapped 95% confidence intervals. Significance was estimated using permutation testing and FDR corrected: *FDR $q < 0.05$, **FDR $q < 0.01$, ***FDR $q < 0.001$. Gray shading indicates the average split-half reliability of repeated presentations of the same videos in the test set for the least reliable subject (lower bound) and most reliable subject (upper bound). EVC, early visual cortex; MT, middle temporal area; EBA, extrastriate body area; LOC, lateral occipital complex; pSTS-SI, social-interaction-selective posterior superior temporal sulcus; STS-Face, face-selective STS; aSTS-SI, anterior STS-SI. ROIs on the surface of individual subjects are shown in Figure S1, and individual subject results are shown in Figure S2.

(B) Whole brain split-half reliability of repeated presentations of the same videos in the test set (related to the gray bars in A).

(C) The whole-brain explained variance by the full encoding model in one example subject (related to the purple bars in A). Whole-brain results in others subjects are shown in Figure S2.

transportation. For this reason—and following the work of others[25]—we differentiate between the spatial configuration of people in a scene and true social interactions. Here, we term these visual cues that may be indicative of the presence of a social interaction "social primitives." Human annotators rated two visual social primitive features: agent distance (how physically far apart the two people in the video are from one another) and "facingness" (the degree to which the people face one another).

Next, we collected ratings for social interaction and affective information in the dataset. Research on social interaction recognition often depicts interactions via coordinated or joint actions between agents (e.g., dancing or chasing[5,26–29]). Some studies also depict communicative actions (e.g., gesturing toward or shaking a fist at someone[4,15,30]), but these studies are in the minority despite the prevalence of communicative actions in daily life. Further, responses to these two different types of interactions have only been compared with a limited extent across studies.[30] For these reasons, we separately collected ratings for joint action and communication (Figures 1B and 1C). We also collected ratings for affective features: valence and arousal.

To ensure that our dataset captured meaningful variance of our annotated features, we visualized the distribution of ratings (Figure 1C). For many of the features, the ratings spanned the full scale of possible values, with the exception of spatial expanse and agent distance.

In addition, to assess the degree of overlap between features, we performed pairwise regressions between each pair of features (Figure 1E). As expected, AlexNet-conv2 predicts many features, most strongly, motion energy and scene and object features and, to a lesser extent, social primitive and social interaction features. Spatial expanse is highly predictive of agent distance. In contrast, social interaction and affective features are not well predicted by scene and object features. Facingness is somewhat predictive of communicative actions.

Although there are correlations present between features, given the naturalistic nature of our stimulus set, further reducing them is neither feasible nor desirable because many features naturally covary in the real world. For example, designing a stimulus set that eliminates the correlation between communication and facingness would yield a set of videos that greatly differ from the statistics of everyday interactions.[31]

Extensive scanning yields high-quality fMRI data

To collect a large amount of high-quality data in individual subjects, we opted for a condition-rich, small-n design.[32,33] Participants (n = 4) viewed the videos in the fMRI scanner over four 2-h scan sessions. Videos were divided into a training (n = 200, presented 9 or 10 times per participant) and test (n = 50, presented 18 or 20 times per participant) set.

In addition to the main experimental runs, participants completed a battery of functional localizers (Figure S1). We localized regions selective for faces (fusiform face area [FFA] and face-selective STS [STS-Face]), bodies (EBA), objects (lateral occipital complex [LOC]), scenes (parahippocampal place area [PPA]), and social interactions (social-interaction-selective posterior STS [pSTS-SI] and social-interaction-selective anterior STS [aSTS-SI]). We additionally defined anatomical ROIs based on Wang et al.[34] (EVC and middle temporal [MT] area). Though we localized ventral regions (FFA and PPA), this paper will focus on responses in lateral regions, given our overall aim of mapping their cortical organization. We present results for these ventral ROIs in the supplement.

To estimate the data quality of the test set, we correlated voxel responses between odd and even presentations of each video. We calculated the average correlation in each of our ROIs and found that the data quality was exceptionally high (mean $r^2 > 0.1$ in every ROI; Figure 2) and thus sufficient for within-subject encoding model analyses. We found particularly high

Please cite this article in press as: McMahon et al., Hierarchical organization of social action features along the lateral visual pathway, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.10.015

reliability in bilateral posterior STS and along the STS in the right hemisphere, which has been absent in prior studies, including others with dynamic, naturalistic stimuli[3] or with much more data for static images.[35] To remove noisy voxels from subsequent encoding model analyses, we used a liberal threshold corresponding to a p value of 0.05 (one-tailed, uncorrected).

Annotated features predict brain responses along the lateral surface of the brain

Before investigating how specific social visual features are organized in the brain, as a first step, we asked whether the combination of all features is predictive of responses in our ROIs. To test this, we fit a voxel-wise encoding model using ordinary least squares (OLS) regression on videos in the training set. From the learned transformation, we predicted responses to videos in the test set. We used the signed-squared correlation between the predicted and actual voxel-wise responses as the prediction metric. After evaluating model performance in every reliable voxel, we calculated the average prediction and estimated variance in prediction for each ROI.

The encoding model predicted responses in nearly every ROI on the lateral and ventral surfaces at the group-level ($r^2 > 0.08$, false discovery rate [FDR] $q < 0.05$, Figure 2A) and individual subject levels (Figure S3). Although prediction in each ROI did not reach the noise ceiling (Figure 2A), our results are similar to other encoding model papers with similarly high-quality, within-subject data.[35–38] These results show that, all together, our features explain significant variance of video responses along the lateral visual pathway. We next sought to investigate how performance varied across different categories of features.

Low- to high-level feature categories predict activity in posterior to anterior ROIs along the lateral surface

To understand the contribution of each group of features to neural responses, we fit a separate encoding model for each of the six categories of features: AlexNet-conv2, motion energy, scene and object, social primitives, social interaction, and affective. This initial feature analysis method allowed us to investigate which regions are predicted by a given feature category without removing shared variance between features.

In low-level regions, we found that responses are largely driven by visual features. In particular, responses in EVC were predicted by AlexNet-conv2, motion energy, and scene and object features ($r^2 > 0.03$, FDR $q < 0.01$). MT was predicted by AlexNet-conv2, motion energy, scene and object, and affective features ($r^2 > 0.02$, FDR $q < 0.05$), although the effect for AlexNet-conv2 in MT was driven by only two subjects (Figures 3A and S3A).

In mid-level regions, EBA and LOC, we found high predictivity of many features. This is likely due to correlations of mid-level features with both low-level visual and higher-level social features in the dataset. EBA was significantly predicted by all feature categories ($r^2 > 0.04$, FDR $q < 0.05$), and LOC was predicted by all categories except social interaction features ($r^2 > 0.03$, FDR $q < 0.05$, Figure 3A).

The three regions in the STS were best predicted by social interaction features. pSTS-SI and STS-Face regions were also predicted by AlexNet-conv2, scene and object, social primitive, and social interaction features ($r^2 > 0.03$, FDR $q < 0.05$). aSTS-SI was predicted by AlexNet-conv2, scene and object, and social features ($r^2 > 0.04$, FDR $q < 0.05$), although the effect for scene and object features was driven by only two of four subjects (Figure 3A).

Within the entire reliability mask, we also visualized which model significantly predicted each voxel's response. If two or more models were significant, we visualized the model that maximally predicted the voxel. In this whole-brain analysis, we also see a progression from low-level features in early areas to abstract features along the lateral surface (Figures 3B and S5A–S5D).

Although many categories are predictive across different regions because of feature correlations in the stimulus set, these results show a progression from the highest predictivity of low-level visual and motion features in lower-level regions (EVC and MT) to social primitives in mid-level regions (EBA and LOC) to social interaction features in regions along the STS (pSTS-SI, STS-Face, and aSTS-SI; Figure 3).

Unique variance is explained by increasingly abstract social features along the lateral surface

To account for shared variance between our feature categories, we next used variance partitioning to calculate the unique variance in voxel responses predicted by each category of features while controlling for variance explained by all other feature categories.

In low-level areas, motion energy uniquely predicted responses in EVC ($r^2 = 0.04$, FDR $q < 0.001$). MT was uniquely predicted by AlexNet-conv2, motion energy, and affective features ($r^2 > 0.05$, FDR $q < 0.05$), although the effect for AlexNet-conv2 features was driven by only two subjects (Figures 4A and S3B).

In mid-level regions (Figure 4A), responses in EBA were uniquely predicted only by motion energy features ($r^2 > 0.02$, FDR $q < 0.01$). Responses in LOC were robustly predicted by scene and object and social primitive features ($r^2 > 0.02$, FDR $q < 0.001$). To a lesser extent, LOC responses were predicted by motion energy and affective features ($r^2 > 0.006$, FDR $q < 0.05$), although the affective feature effect is driven by only two subjects (Figure S3B).

Along the STS (Figure 4A), all ROIs (pSTS-SI, STS-Face, aSTS-SI) were uniquely predicted only by social interaction features ($r^2 > 0.03$, FDR $q < 0.01$), although the effect in pSTS was driven by only two out of four subjects.

These results further reveal a pattern of increasing abstractness of features along the lateral surface: low-level visual and motion features explain unique variance in EVC and MT, scene and object and social primitive features in LOC, and social interaction features along the STS (particularly in the two more anterior regions). This gradient can also be observed in the whole brain predictivity on the lateral surface (Figures 4B and S5E–S5H). Although there is some variability across subjects at the whole brain level, each subject shows a posterior to anterior gradient of low- (AlexNet-conv2 and motion energy) to mid- (scene and object and social primitive) to high-level social interaction features.

Communication uniquely predicts STS responses

In our final, and most stringent, analysis, we moved beyond feature categories to investigate where individual features in our stimuli explained unique variance relative to all other features, including those in the same feature category.

In low-level regions, some features predicted responses in one or two individual subjects, but no single feature robustly

Please cite this article in press as: McMahon et al., Hierarchical organization of social action features along the lateral visual pathway, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.10.015

CellPress
OPEN ACCESS

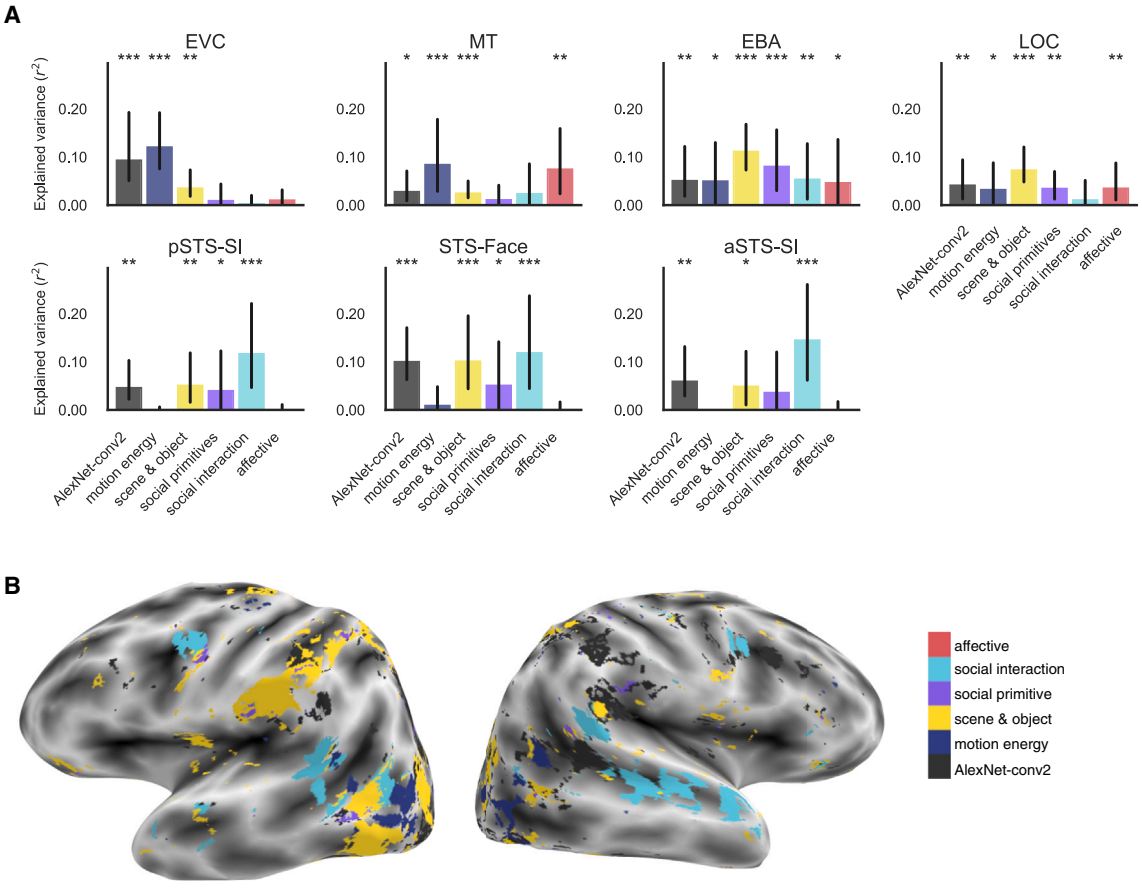

Figure 3. Low- to high-level feature categories predict activity along the lateral surface

(A) Group-level explained variance by each feature category in each of the lateral ROIs. Individual subject results are shown in Figure S3A and ventral ROI results in Figure S4A–S4B. Error bars are the bootstrapped 95% confidence intervals. Significance was estimated using permutation testing and FDR corrected: *FDR $q < 0.05$, **FDR $q < 0.01$, ***FDR $q < 0.001$.

(B) Encoding model preference map in one example subject. Voxels are colored by the category with the greatest significant prediction following multiple comparison correction. Other subjects are in Figures S5A–S5D.

predicted EVC or MT responses. In particular, some small but significant effects were observed at the group level for communication, valence, and arousal in EVC or MT, but these effects were driven by only one or two subjects each ($r^2 > 0.001$, FDR $q < 0.05$; Figures 5A and S3C).

In mid-level regions (Figure 5A), unique variance in EBA was not predicted by any feature ($r^2 < 0.007$, FDR $q > 0.05$), but LOC was uniquely predicted by several mid-level features, including object directedness, agent distance (in only two subjects), and facingness ($r^2 > 0.005$, FDR $q < 0.05$).

Along the STS (Figure 5A), all regions were uniquely predicted by communication ($r^2 > 0.03$, FDR $q < 0.01$), although the effect in pSTS-SI was driven by only two subjects. Surprisingly, the magnitude of these effects was the same as both combined social interaction features, suggesting the joint action feature explains little variance along the STS. At the group-level, aSTS-SI was also predicted by spatial expanse ($r^2 = 0.007$, FDR $q < 0.05$), but this effect was only present in one subject.

For object directedness and communication, which robustly predicted responses at the group and individual subject level, we visualized how well the features uniquely predicted responses within all cortical voxels in the reliability mask (communication: Figures 5B and S6E–S6H; object directedness: Figures S6A–S6D). Though robust in the ROI analysis, facingness only survived multiple comparisons correction in the whole brain in two out of four subjects. We find representations of object directedness most strongly in lateral occipital regions of the left hemisphere (Figures S6A–S6D), and communication represented along the STS most strongly in the right hemisphere (Figures S6E–S6H). These results further confirm the pattern of increasing abstractness along the lateral surface of the brain and, moreover, provide strong evidence that communicative actions particularly drive responses in the STS.

Social interaction selectivity along the STS is not driven by face size or position

In addition to uniquely predicting responses along the STS (Figure 4A), social interaction features also uniquely predict activity in the FFA ($r^2 = 0.02$, FDR $q < 0.01$; Figure S4C). This raises the concern that social interaction features may be confounded with face features (e.g., face size or position). However, the pattern of responses in STS regions and FFA appear to be quite

Please cite this article in press as: McMahon et al., Hierarchical organization of social action features along the lateral visual pathway, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.10.015

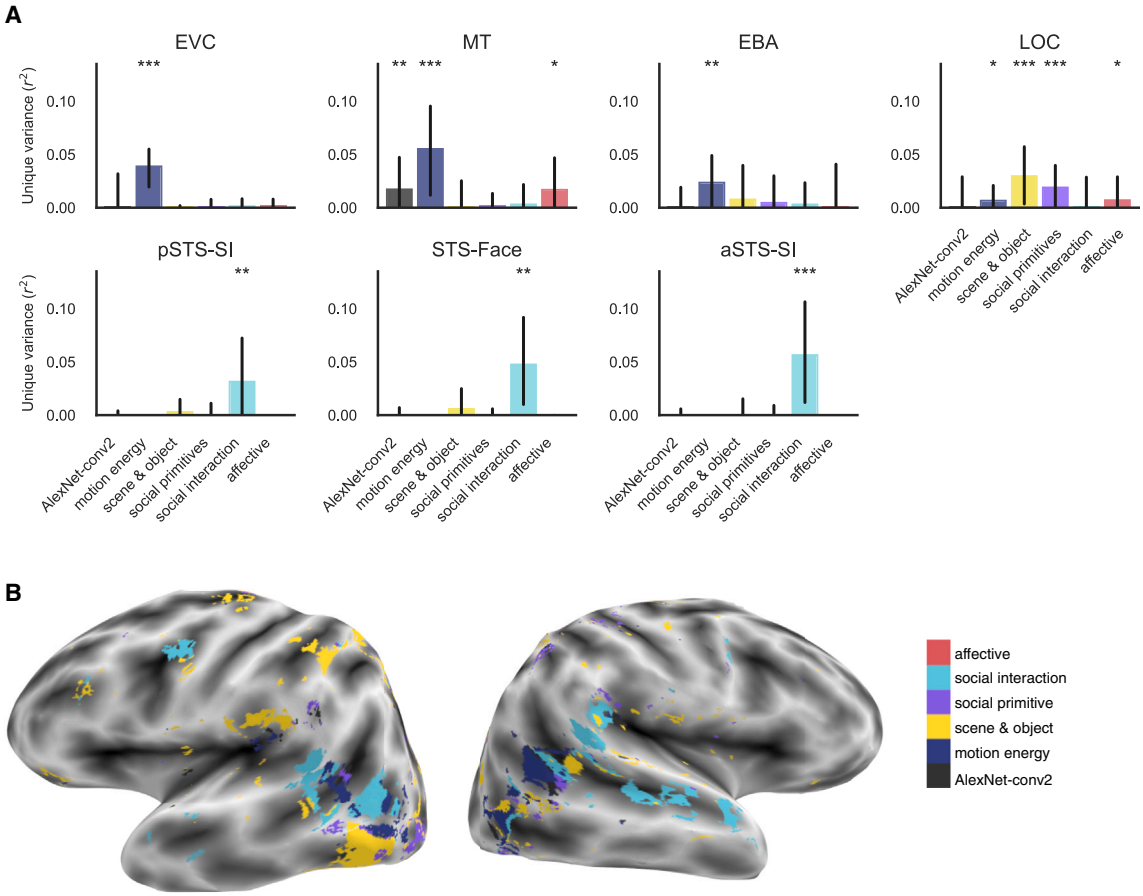

Figure 4. Unique variance is explained by increasingly abstract social features along the lateral surface

(A) Group-level unique variance explained by each feature category in each of the lateral ROIs. Individual subject results are shown in Figure S3B and ventral ROI results in Figure S4C–S4D. Error bars are the bootstrapped 95% confidence intervals. Significance was estimated using permutation testing and FDR corrected: *FDR $q < 0.05$, **FDR $q < 0.01$, ***FDR $q < 0.001$.

(B) Encoding model preference map in one example subject. Voxels are colored by the category with the greatest significant unique prediction following multiple comparison correction. Other subjects are in Figures S5E–S5H.

different—social interaction features are more predictive than other categories only in STS regions. To test this directly, we performed a non-parametric ANOVA comparing ROI (pSTS-SI or aSTS-SI versus FFA) and category (social interaction features versus scene and object features) and found a significantly greater relative response to social interactions in both pSTS-SI and aSTS-SI in all four subjects ($r^2 > 0.02$, FDR $q < 0.05$). These results suggest that social interaction representations in the STS are distinct from those in the FFA and unlikely to be the result of face confounds in the stimulus set.

To further ensure that our social feature annotations were not simply a product of the size and position of faces in the videos, we annotated the location of faces in the scene using bounding boxes and quantified the size and position of the faces as face area and face centrality. Face area is the sum of the area of the two face-bounding boxes averaged across frames, and face centrality is the minimum distance of the two bounding boxes from the center of the frame averaged across frames.

We found that face area was related to all scene and object and social primitive features ($r > \pm 0.15$, FDR $q < 0.05$) but not social interaction or affective features ($r < 0.13$, FDR $q > 0.05$; Figure 6A). Face centrality was only correlated with the social primitive features, agent distance and facingness ($r > 0.21$, FDR $q < 0.004$; Figure 6B). Because face size and position were only correlated with descriptors of the scene or configuration of people within the scene, this demonstrates that our communication feature is not confounded with the size and position of faces in the videos.

Social interaction selectivity is not explained by looking at faces

In a post hoc analysis, we investigated whether the results presented here may be the result of how people looked at the videos. Two fMRI participants returned to participate in an eye tracking experiment (sub-02 and sub-03) in addition to a group of new participants (n = 11 in total). All participants viewed the 50 test set videos while their gaze was tracked. From the eye tracking, we computed a heatmap of fixations (Figure 6C). From the heatmaps, we found that participants viewed the videos in a highly consistent manner, both across presentations within participants (average within-subject split-half reliability = 0.51, range = 0.39–0.63, sub-02 = 0.62, sub-03 = 0.48) and between participants (average

Please cite this article in press as: McMahon et al., Hierarchical organization of social action features along the lateral visual pathway, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.10.015

CellPress
OPEN ACCESS

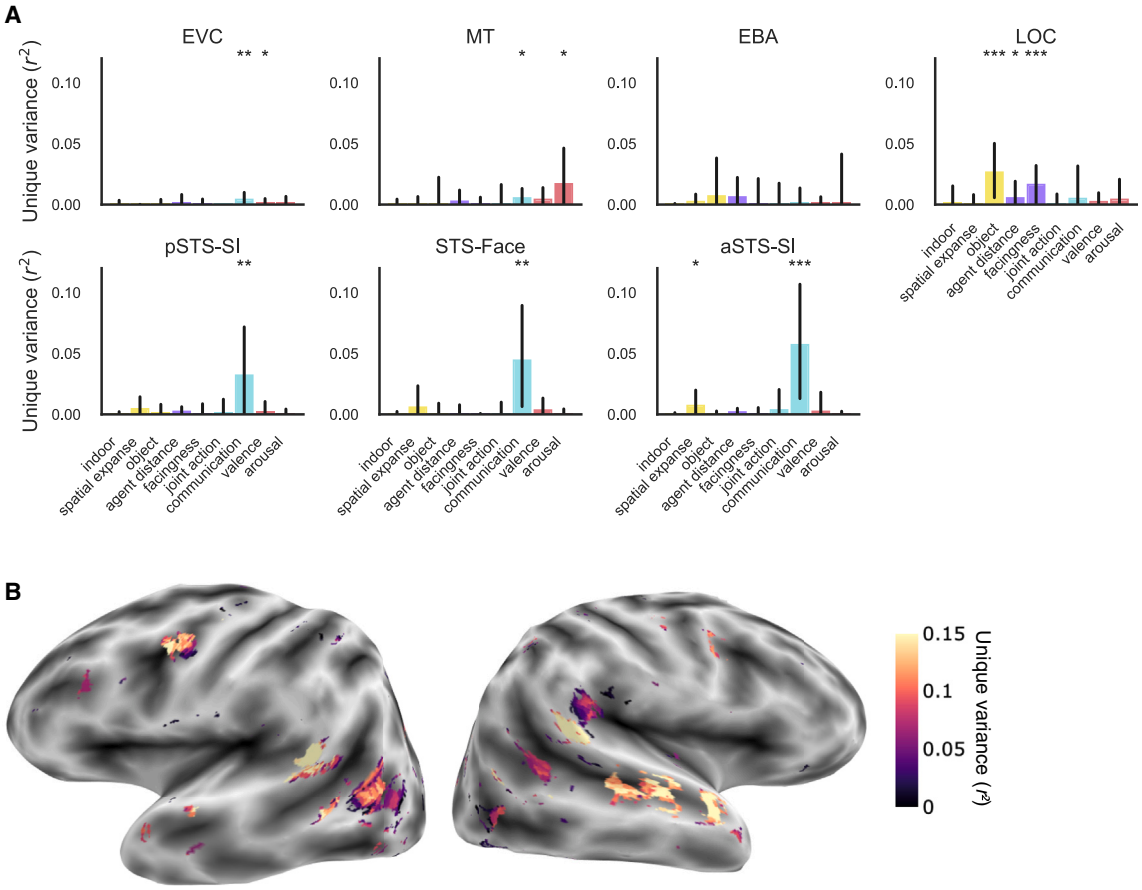

Figure 5. Communication uniquely predicts STS responses

(A) Group-level unique variance explained by each individual annotated feature in each of the lateral ROIs. Individual subject results are shown in Figure S3C and ventral ROI results in Figure S4E–S4F. Error bars are the bootstrapped 95% confidence intervals. Significance was estimated using permutation testing and FDR corrected: *FDR $q < 0.05$, **FDR $q < 0.01$, ***FDR $q < 0.001$.

(B) Unique variance explained by the communication feature in one example subject. Map is thresholded to FDR $q < 0.05$. Other subjects are shown in Figure S6.

inter-subject reliability = 0.59, range = 0.41–0.72). The slightly higher inter-subject than within-subject reliability is likely the result of a less noisy estimate due to averaging over more trials and participants (see STAR Methods).

The high inter-subject reliability validates the approach of investigating the gaze pattern in a new group of participants. More importantly, the high within-subject reliability suggests that the pattern of fixation was highly similar across repetitions. We investigated whether participants looked more consistently at some types of videos than others by correlating both the within- and between-subject reliability with annotated features. We found no relation between any annotated feature and the consistency of viewing within ($r < \pm0.13$, FDR $q > 0.05$) or between subjects ($r < \pm0.17$, FDR $q > 0.05$).

People tend to look at faces,[39–45] which is true for our videos as well (average proportion of samples within face-bounding boxes = 0.4, range = 0.26–0.47). As a result of this finding, we investigated whether the proportion of time that participants spent looking at faces was related to the content of the videos. We found that most annotated features, except object directedness and facingness ($r < \pm0.06$, FDR $q > 0.05$), were correlated with the proportion of time that participants spent looking at faces ($r > \pm0.13$, FDR $q < 0.001$; Figure 6D). Spatial expanse was significantly more related to the proportion of looking at faces than communication (difference of absolute correlation = 0.26, $p < 0.001$). Thus, the time spent looking at faces is more strongly related to scene content than social content of the videos. This is most likely because faces tend to appear larger in close-up scenes and smaller in larger scenes.

Together, the eye tracking results strongly suggest that the finding of social interaction selectivity, and communication selectivity in particular, is not explainable by participants' viewing behavior.

DISCUSSION

Here, we investigated the organization of features of social actions on the lateral surface of the human brain. We introduced a rich video dataset and accompanying high-quality fMRI data to study social actions in naturalistic contexts. Using voxel-wise encoding models and variance partitioning, we found that social features are organized from low-level

Please cite this article in press as: McMahon et al., Hierarchical organization of social action features along the lateral visual pathway, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.10.015

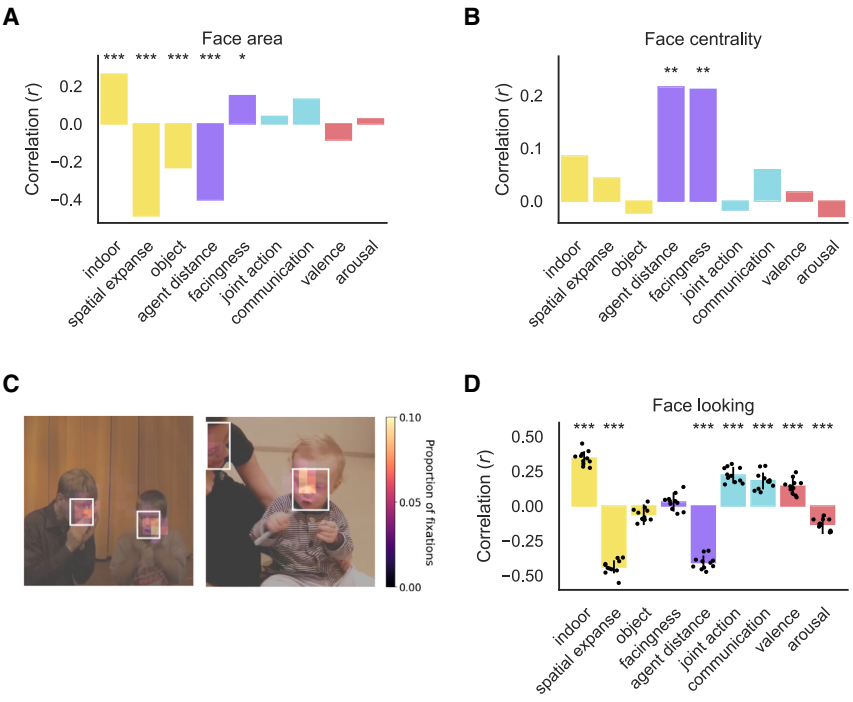

Figure 6. Social interaction selectivity is not explained by looking at faces

(A and B) The correlation between face area (A) and face centrality (B) with feature ratings across the full stimulus set. Significant correlations determined through permutation test are marked: *FDR $q < 0.05$, **FDR $q < 0.01$, ***FDR $q < 0.001$.

(C) Average heatmap across participants overlaid on still frames from two example videos. The white boxes indicated the annotated face-bounding boxes.

(D) The correlation between the proportion of time spent looking at faces on each video correlated with each of the annotated features. Each black dot is the correlation for each eye tracking participant. Error bars are the bootstrapped 95% confidence interval (CI) averaged across participants. Significant correlations are marked as in (A) and (B).

features in EVC and MT, to mid-level features in EBA and LOC, and abstract social features in the STS. Surprisingly, we found that communicative actions in particular drive responses along the STS.

Organization of the lateral stream

In their opinion piece arguing for a third visual stream specialized for social perception, Pitcher and Ungerleider[8] suggested that the lateral stream is organized hierarchically, with projections from the EVC to the MT and STS, and computes a range of functions based on dynamic social cues. By broadly sampling the space of dynamic, social actions, we provide the first strong empirical test of hierarchical processing of increasingly abstract features along the lateral stream. This is consistent with other work that suggested similar organization based on a smaller number of features.[2,46] The combination of rich, high-quality data and comprehensive feature sampling features allows us to unify and expand findings across these prior studies.

Wurm and Caramazza[7] also recently argued that the lateral surface is involved in social processing but that the lateral occipitotemporal cortex (LOTC) is specialized for action perception in particular. They present evidence for an object-to-person-directed, ventral-to-dorsal organization of action-relevant features on the lateral surface, similar to the broad inanimate-to-animate organization of ventral occipitotemporal cortex. Investigating animacy was not the main focus of our current study. However, we do find that scene and object features are represented in a more ventral region of the LOTC (i.e., object-selective LOC), and person-related features, like social primitive and communicative actions, are represented in more dorsal regions (i.e., EBA and STS). However, we also find social primitive representations in the LOC. Further, we find preliminary evidence of communicative representations most strongly in the right hemisphere and object responses in the left hemisphere (Figure S6), while their proposal centers solely on the left hemisphere.[2,7] Generally, we interpret this organizational structure as evidence for hierarchical processing in which intermediate representations in the LOC support the computational goal of higher-level regions along the STS.

Communicative actions

We find evidence that representing communicative action may be a key computational goal of the lateral visual stream, but it may be one goal among others. For instance, the STS has been implicated in many different functions, including dynamic face perception, biological motion perception, speech processing, theory of mind, and audiovisual integration.[47,48]

Previous work has established that there are selective responses to the presence of social interactions in a region of the pSTS in both controlled[4,6,15] and naturalistic[13] stimuli. Some of these studies used communicative actions[4,13,15] and others used Heider and Simmel[49] displays that, while not communicative, have narrative structure.[4,6,14] None of these neuroimaging studies investigated differential responses to joint versus communicative actions, but behavioral work with controlled stimuli suggests that both communicative and joint actions are processed preferentially.[28,30] Here, we find robust unique representations of communicative, but not joint, actions along the STS. This finding suggests that STS responses to social interactions are, more specifically, driven by communicative actions.

Although responses to communicative actions in the STS have been shown in prior studies,[50,51] these studies all focused on *second-person* or participant-directed communicative responses. In fact, prior work has argued that the goal of the STS is to process such participant-directed interactions.[52] Here, we find that *observed third-party* communicative interactions strongly and robustly drive responses in regions along the STS. This finding builds significantly on prior work by suggesting that the STS is not only involved in processing the communicative actions of one's own social partners but also processes all observed communicative actions. As we did not

Please cite this article in press as: McMahon et al., Hierarchical organization of social action features along the lateral visual pathway, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.10.015

include any participant-directed videos, it remains an open question to what extent representations of communicative actions directed at oneself versus others overlap in the STS.

The representation of communication here is not driven by hearing speech or language, as our videos were shown to participants without sound. It is possible, however, that the communication results are driven by "visual speech" (e.g., seeing mouth movements[47,51,53–55]) rather than a more abstract representation of communication. This seems unlikely for a few reasons. First, the size and position of faces in the scene was unrelated to the presence of communication (Figures 6A and 6B). Further, while there is a trend to look more at faces for communicative videos, the time spent looking at faces is more related to the spatial expanse of the scene (likely due to the relation between spatial expanse and the size of faces in the videos; Figures 6A and 6D). Finally, the communication responses, particularly in more anterior regions, are largely right lateralized (Figure S6) suggesting that they are not the result of simulated speech or language (both of which are largely left lateralized[56]).

Joint actions

A large body of behavioral work suggests that joint actions are processed in a preferential manner.[26–29,57–59] However, we do not find evidence that joint actions are uniquely represented in the STS or elsewhere. It may be the case that joint action is confounded with other features in the stimulus set, but given the number of features in our dataset, it is infeasible to investigate the shared variance between every combination of features. Future studies with controlled stimuli could better disentangle communicative and joint actions in order to investigate whether joint actions drive responses in the STS—but possibly to a lesser extent than communicative actions. An alternative possibility is that the type of social interactions that the human mind and brain particularly cares about are communicative actions.

Affective features

Although affective features are thought to be represented largely in subcortical regions, some recent work has found affective representations of sensory content in ventral visual regions.[60] Here, we investigated whether valence and arousal are also represented in lateral regions. Previous research found that the goal compatibility of agents (helping versus hindering or cooperation versus competition) is represented in the pSTS[4,6] suggesting that this region may represent the affective content of the social interactions (e.g., positive versus negative interactions). Although we did find representations of affective features in early- and mid-level lateral regions (MT, EBA, and LOC) in our most permissive analysis, we did not find evidence that affective features explain unique variance along the lateral stream. Future studies that better deconfound affective features from visual and social features may be better able to answer this question.

Social primitive features

As in previous work,[10,11,15] in our most permissive analysis, we found that the EBA represents social primitive features, such as whether two people were facing one another and their spatial distance. However, after controlling for other features, we did not find unique representations of social primitives in the EBA although we did find unique variance explained in the nearby LOC. Prior studies have found selectivity for facing bodies outside of the EBA in nearby regions, but they did not localize the LOC, so it is not possible to know whether their activations fall within the object-selective cortex.[10,11] It is important to note that the EBA and LOC are extremely close by and overlapping in most individual subjects (Figure S1). Why may object-selective cortex represent the facingness and distance of bodies? One possibility is that the LOC represents relational information about objects more generally.[61]

In the current study, we adopted the perspective that the configuration of bodies can be a visual cue indicative of a social interaction, but that these cues do not make a social interaction in and of themselves. This has been argued for by others,[25] though controlled stimuli such as nearby facing bodies are often referred to as social interactions in other work.[12,61,62] Here, we find evidence that social primitives are represented in mid-level regions in the lateral stream and more abstract social interaction features in higher-level areas. Our results are consistent with the idea that the visual system uses cues like distance and facingness as precursors to process social interactions.

Motion features

In this study, while our motion energy features predict responses in area MT, they are not predictive of responses in regions along the STS, even in our most permissive analysis (Figure 3A). This may seem in contrast with prior findings[63,64] of strong responses in the STS when viewing or attending to agentic motion. However, one of these studies[63] found that motion was less predictive of fMRI responses when the videos did not contain social features. In our case, we did not separately model motion depending on the source (e.g., object, camera, and social motion). In addition, our motion features are relatively low-level and thus unlikely to capture the more complicated motion patterns that distinguish between biological and non-biological motion. With this in mind, and based on strong evidence that dynamic stimuli drive STS responses to a much greater extent than static stimuli in both humans[64] and macaques,[63] and for different kinds of social content,[65] we do not take our current finding as evidence that the STS does not respond to motion but instead, as consistent with the hypothesis that the STS is selective for agentive motion, in particular.

Beyond the lateral stream

In ventral regions, the primary group of features we see represented are scene and object features. It is unsurprising that scene features, particularly indoor scenes and spatial expanse, are represented in the PPA. Scene features in our dataset are also heavily confounded with the size and visibility of faces in our videos (e.g., close-up videos have larger and more clearly visible faces; Figure 6).

Although we only localized regions within the ventral and lateral visual streams, we also see reliable responses to our stimuli in parietal and frontal regions, which are sometimes referred to as part of the action observation network[66] and have also been implicated in social interaction recognition.[67] However, we do not find any feature representations in these regions that are consistent across subjects (Figure S5). Thus, while these regions are responding consistently to the videos in our dataset, they are not representing the features related to social action considered here. While some prior work has found functional responses to faces in the inferior

frontal gyrus (IFG) and STS to be similar,[68,69] suggesting that IFG may also represent social content, face- and social-interaction-selective regions in the STS, although nearby, are disassociated.[4] . Thus, our results are consistent with growing evidence that these regions do not represent social features of actions.[7]

Concluding remarks

Here, we find evidence for increasingly abstract social feature representations along the lateral visual stream. We also find that regions along the STS are particularly responsive to communicative actions. Understanding how these brain responses relate to those specialized for other types of communicative signals, particularly via language, opens exciting avenues for future research.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Video annotation participants
  - fMRI participants
  - Eye tracking participants
- METHOD DETAILS
  - Stimulus set
  - fMRI Experiment
  - Eye tracking experiment
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Feature regression
  - fMRI Preprocessing
  - fMRI GLM
  - ROI definition in native space
  - Voxel-wise encoding models
  - Evaluating face-related effects
  - Eye tracking

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2023.10.015.

ACKNOWLEDGMENTS

This work was funded in part by NSF GRFP DGE-1746891 awarded to E.M. and NIMH R01MH132826 awarded to L.I. We would like to thank Elahé Yargholi and Maryam Vaziri-Pashkam for sharing their biological motion localization stimuli and scripts. Thank you to Diana Dima, Manasi Malik, and Raj Magesh Gauthaman for helpful comments on earlier versions of this paper.

AUTHOR CONTRIBUTIONS

All authors developed the concept for the paper and planned the experiments. E.M. performed the experiments, analyzed the data, wrote the manuscript, and designed the figures. All authors contributed to finalizing the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: April 26, 2023
Revised: September 1, 2023
Accepted: October 10, 2023
Published: November 1, 2023

REFERENCES

1. Dima, D.C., Tomita, T.M., Honey, C.J., and Isik, L. (2022). Social-affective features drive human representations of observed actions. eLife *11*, e75027. https://doi.org/10.7554/eLife.75027.
2. Wurm, M.F., Caramazza, A., and Lingnau, A. (2017). Action categories in lateral occipitotemporal cortex are organized along sociality and transitivity. J. Neurosci. *37*, 562–575. https://doi.org/10.1523/JNEUROSCI.1717-16.2016.
3. Tarhan, L., and Konkle, T. (2020). Sociality and interaction envelope organize visual action representations. Nat. Commun. *11*, 3002. https://doi.org/10.1038/s41467-020-16846-w.
4. Isik, L., Koldewyn, K., Beeler, D., and Kanwisher, N. (2017). Perceiving social interactions in the posterior superior temporal sulcus. Proc. Natl. Acad. Sci. USA *114*, E9145–E9152. https://doi.org/10.1073/pnas.1714471114.
5. Sliwa, J., and Freiwald, W.A. (2017). A dedicated network for social interaction processing in the primate brain. Science *356*, 745–749. https://doi.org/10.1126/science.aam6383.
6. Walbrin, J., Downing, P., and Koldewyn, K. (2018). Neural responses to visually observed social interactions. Neuropsychologia *112*, 31–39. https://doi.org/10.1016/j.neuropsychologia.2018.02.023.
7. Wurm, M.F., and Caramazza, A. (2022). Two 'what' pathways for action and object recognition. Trends Cogn. Sci. *26*, 103–116. https://doi.org/10.1016/j.tics.2021.10.003.
8. Pitcher, D., and Ungerleider, L.G. (2021). Evidence for a third visual pathway specialized for social perception. Trends Cogn. Sci. *25*, 100–110. https://doi.org/10.1016/j.tics.2020.11.006.
9. DiCarlo, J.J., Zoccolan, D., and Rust, N.C. (2012). How does the brain solve visual object recognition? Neuron *73*, 415–434. https://doi.org/10.1016/j.neuron.2012.01.010.
10. Abassi, E., and Papeo, L. (2020). The representation of two-body shapes in the human visual cortex. J. Neurosci. *40*, 852–863. https://doi.org/10.1523/JNEUROSCI.1378-19.2019.
11. Abassi, E., and Papeo, L. (2022). Behavioral and neural markers of visual configural processing in social scene perception. NeuroImage *260*, 119506. https://doi.org/10.1016/j.neuroimage.2022.119506.
12. Landsiedel, J., Daughters, K., Downing, P.E., and Koldewyn, K. (2022). The role of motion in the neural representation of social interactions in the posterior temporal cortex. NeuroImage *262*, 119533. https://doi.org/10.1016/j.neuroimage.2022.119533.
13. Lee Masson, H., and Isik, L. (2021). Functional selectivity for social interaction perception in the human superior temporal sulcus during natural viewing. NeuroImage *245*, 118741. https://doi.org/10.1016/j.neuroimage.2021.118741.
14. Varrier, R.S., and Finn, E.S. (2022). Seeing social: A neural signature for conscious perception of social interactions. J. Neurosci. *42*, 9211–9226. https://doi.org/10.1523/JNEUROSCI.0859-22.2022.
15. Walbrin, J., and Koldewyn, K. (2019). Dyadic interaction processing in the posterior temporal cortex. NeuroImage *198*, 296–302. https://doi.org/10.1016/j.neuroimage.2019.05.027.
16. Redcay, E., and Moraczewski, D. (2020). Social cognition in context: a naturalistic imaging approach. NeuroImage *216*, 116392. https://doi.org/10.1016/j.neuroimage.2019.116392.

17. Haxby, J.V., Gobbini, M.I., and Nastase, S.A. (2020). Naturalistic stimuli reveal a dominant role for agentic action in visual representation. NeuroImage *216*, 116561. https://doi.org/10.1016/j.neuroimage.2020.116561.

18. Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S.A., Yan, T., Brown, L., Fan, Q., Gutfruend, D., Vondrick, C., et al. (2019). Moments in time dataset: one million videos for event understanding. IEEE Trans. Pattern Anal. Mach. Intell. *42*, 502–508. https://doi.org/10.1109/TPAMI.2019.2901464.

19. U.S. Bureau of Labor Statistics; U.S. Census Bureau (2019). American Time Use Survey — 2019 Results. https://www.bls.gov/news.release/archives/atus_06252020.pdf.

20. Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). ImageNet classification with deep convolutional neural networks. In Adv. Neural Inf. Process. Syst., *25*, F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, eds. (Curran Associates, Inc.), pp. 1097–1105.

21. Adelson, E.H., and Bergen, J.R. (1985). Spatiotemporal energy models for the perception of motion. J. Opt. Soc. Am. A *2*, 284–299. https://doi.org/10.1364/JOSAA.2.000284.

22. Tucciarelli, R., Wurm, M., Baccolo, E., and Lingnau, A. (2019). The representational space of observed actions. eLife *8*, e47686. https://doi.org/10.7554/eLife.47686.

23. Papeo, L. (2020). Twos in human visual perception. Cortex *132*, 473–478. https://doi.org/10.1016/j.cortex.2020.06.005.

24. Zhou, C., Han, M., Liang, Q., Hu, Y.-F., and Kuai, S.-G. (2019). A social interaction field model accurately identifies static and dynamic social groupings. Nat. Hum. Behav. *3*, 847–855. https://doi.org/10.1038/s41562-019-0618-2.

25. Hochmann, J.-R., and Papeo, L. (2021). How can it be both abstract and perceptual? Comment on Hafri, A., & Firestone, C. (2021), The perception of relations, Trends in Cognitive Sciences. https://doi.org/10.31234/osf.io/hm49p.

26. Gao, T., Newman, G.E., and Scholl, B.J. (2009). The psychophysics of chasing: A case study in the perception of animacy. Cogn. Psychol. *59*, 154–179. https://doi.org/10.1016/j.cogpsych.2009.03.001.

27. Gao, T., McCarthy, G., and Scholl, B.J. (2010). The wolfpack effect: perception of animacy irresistibly influences interactive behavior. Psychol. Sci. *21*, 1845–1853. https://doi.org/10.1177/0956797610388814.

28. Neri, P., Luu, J.Y., and Levi, D.M. (2006). Meaningful interactions can enhance visual discrimination of human agents. Nat. Neurosci. *9*, 1186–1192. https://doi.org/10.1038/nn1759.

29. Quadflieg, S., and Koldewyn, K. (2017). The neuroscience of people watching: how the human brain makes sense of other people's encounters. Ann. N. Y. Acad. Sci. *1396*, 166–182. https://doi.org/10.1111/nyas.13331.

30. Manera, V., Del Giudice, M., Bara, B.G., Verfaillie, K., and Becchio, C. (2011). The second-agent effect: communicative gestures increase the likelihood of perceiving a second agent. PLoS One *6*, e22650. https://doi.org/10.1371/journal.pone.0022650.

31. Grall, C., and Finn, E.S. (2022). Leveraging the power of media to drive cognition: a media-informed approach to naturalistic neuroscience. Soc. Cogn. Affect. Neurosci. *17*, 598–608. https://doi.org/10.1093/scan/nsac019.

32. Mahowald, K., and Fedorenko, E. (2016). Reliable individual-level neural markers of high-level language processing: A necessary precursor for relating neural variability to behavioral and genetic variability. NeuroImage *139*, 74–93. https://doi.org/10.1016/j.neuroimage.2016.05.073.

33. Naselaris, T., Allen, E., and Kay, K. (2021). Extensive sampling for complete models of individual brains. Curr. Opin. Behav. Sci. *40*, 45–51. https://doi.org/10.1016/j.cobeha.2020.12.008.

34. Wang, L., Mruczek, R.E.B., Arcaro, M.J., and Kastner, S. (2015). Probabilistic maps of visual topography in human cortex. Cereb. Cortex *25*, 3911–3931. https://doi.org/10.1093/cercor/bhu277.

35. Allen, E.J., St-Yves, G., Wu, Y., Breedlove, J.L., Prince, J.S., Dowdle, L.T., Nau, M., Caron, B., Pestilli, F., Charest, I., et al. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. Nat. Neurosci. *25*, 116–126. https://doi.org/10.1038/s41593-021-00962-x.

36. Bonner, M.F., and Epstein, R.A. (2018). Computational mechanisms underlying cortical responses to the affordance properties of visual scenes. PLoS Comput. Biol. *14*, e1006111. https://doi.org/10.1371/journal.pcbi.1006111.

37. Çukur, T., Nishimoto, S., Huth, A.G., and Gallant, J.L. (2013). Attention during natural vision warps semantic representation across the human brain. Nat. Neurosci. *16*, 763–770. https://doi.org/10.1038/nn.3381.

38. Huth, A.G., Nishimoto, S., Vu, A.T., and Gallant, J.L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron *76*, 1210–1224. https://doi.org/10.1016/j.neuron.2012.10.014.

39. Bindemann, M., Burton, A.M., Hooge, I.T.C., Jenkins, R., and de Haan, E.H.F. (2005). Faces retain attention. Psychon. Bull. Rev. *12*, 1048–1053. https://doi.org/10.3758/BF03206442.

40. Gamer, M., and Büchel, C. (2009). Amygdala activation predicts gaze toward fearful eyes. J. Neurosci. *29*, 9123–9126. https://doi.org/10.1523/JNEUROSCI.1883-09.2009.

41. Mack, A., Pappas, Z., Silverman, M., and Gay, R. (2002). What we see: inattention and the capture of attention by meaning. Conscious. Cogn. *11*, 488–506. https://doi.org/10.1016/S1053-8100(02)00028-4.

42. Ro, T., Russell, C., and Lavie, N. (2001). Changing faces: A detection advantage in the flicker paradigm. Psychol. Sci. *12*, 94–99. https://doi.org/10.1111/1467-9280.00317.

43. Shelley-Tremblay, J., and Mack, A. (1999). Metacontrast masking and attention. Psychol. Sci. *10*, 508–515. https://doi.org/10.1111/1467-9280.00197.

44. Theeuwes, J., and Van der Stigchel, S. (2006). Faces capture attention: evidence from inhibition of return. Vis. Cogn. *13*, 657–665. https://doi.org/10.1080/13506280500410949.

45. Vuilleumier, P. (2000). Faces call for attention: evidence from patients with visual extinction. Neuropsychologia *38*, 693–700. https://doi.org/10.1016/S0028-3932(99)00107-4.

46. Tarhan, L., De Freitas, J., and Konkle, T. (2021). Behavioral and neural representations en route to intuitive action understanding. Neuropsychologia *163*, 108048. https://doi.org/10.1016/j.neuropsychologia.2021.108048.

47. Deen, B., Koldewyn, K., Kanwisher, N., and Saxe, R. (2015). Functional organization of social perception and cognition in the superior temporal sulcus. Cereb. Cortex *25*, 4596–4609. https://doi.org/10.1093/cercor/bhv111.

48. Hein, G., and Knight, R.T. (2008). Superior temporal sulcus–It's my area: or is it? J. Cogn. Neurosci. *20*, 2125–2136. https://doi.org/10.1162/jocn.2008.20148.

49. Heider, F., and Simmel, M. (1944). An experimental study of apparent behavior. Am. J. Psychol. *57*, 243–259. https://doi.org/10.2307/1416950.

50. Redcay, E., Velnoskey, K.R., and Rowe, M.L. (2016). Perceived communicative intent in gesture and language modulates the superior temporal sulcus. Hum. Brain Mapp. *37*, 3444–3461. https://doi.org/10.1002/hbm.23251.

51. Deen, B., Saxe, R., and Kanwisher, N. (2020). Processing communicative facial and vocal cues in the superior temporal sulcus. NeuroImage *221*, 117191. https://doi.org/10.1016/j.neuroimage.2020.117191.

52. Redcay, E. (2008). The superior temporal sulcus performs a common function for social and speech perception: implications for the emergence of autism. Neurosci. Biobehav. Rev. *32*, 123–142. https://doi.org/10.1016/j.neubiorev.2007.06.004.

53. Callan, D.E., Jones, J.A., Munhall, K., Kroos, C., Callan, A.M., and Vatikiotis-Bateson, E. (2004). Multisensory integration sites identified by perception of spatial wavelet filtered visual speech gesture information. J. Cogn. Neurosci. *16*, 805–816. https://doi.org/10.1162/089892904970771.

54. Calvert, G.A., Bullmore, E.T., Brammer, M.J., Campbell, R., Williams, S.C.R., McGuire, P.K., Woodruff, P.W.R., Iversen, S.D., and David, A.S.

(1997). Activation of auditory cortex during silent lipreading. Science *276*, 593–596. https://doi.org/10.1126/science.276.5312.593.

55. Capek, C.M., MacSweeney, M., Woll, B., Waters, D., McGuire, P.K., David, A.S., Brammer, M.J., and Campbell, R. (2008). Cortical circuits for silent speechreading in deaf and hearing people. Neuropsychologia *46*, 1233–1241. https://doi.org/10.1016/j.neuropsychologia.2007.11.026.
56. Malik-Moraleda, S., Ayyash, D., Gallée, J., Affourtit, J., Hoffmann, M., Mineroff, Z., Jouravlev, O., and Fedorenko, E. (2022). An investigation across 45 languages and 12 language families reveals a universal language network. Nat. Neurosci. *25*, 1014–1019. https://doi.org/10.1038/s41593-022-01114-5.
57. Fedorov, L.A., Chang, D.S., Giese, M.A., Bülthoff, H.H., and de la Rosa, S. (2018). Adaptation aftereffects reveal representations for encoding of contingent social actions. Proc. Natl. Acad. Sci. USA *115*, 7515–7520. https://doi.org/10.1073/pnas.1801364115.
58. Hafri, A., Papafragou, A., and Trueswell, J.C. (2013). Getting the gist of events: recognition of two-participant actions from brief displays. J. Exp. Psychol. Gen. *142*, 880–905. https://doi.org/10.1037/a0030045.
59. Hafri, A., Trueswell, J.C., and Strickland, B. (2018). Encoding of event roles from visual scenes is rapid, spontaneous, and interacts with higher-level visual processing. Cognition *175*, 36–52. https://doi.org/10.1016/j.cognition.2018.02.011.
60. Kragel, P.A., Reddan, M.C., LaBar, K.S., and Wager, T.D. (2019). Emotion schemas are embedded in the human visual system. Sci. Adv. *5*, eaaw4358. https://doi.org/10.1126/sciadv.aaw4358.
61. Hafri, A., and Firestone, C. (2021). The perception of relations. Trends Cogn. Sci. *25*, 475–492.
62. Malik, M., and Isik, L. (2022). Relational visual information explains human social inference: a graph neural network model for social interaction recognition. https://doi.org/10.31234/osf.io/5cuyr.
63. Russ, B.E., and Leopold, D.A. (2015). Functional MRI mapping of dynamic visual features during natural viewing in the macaque. NeuroImage *109*, 84–94. https://doi.org/10.1016/j.neuroimage.2015.01.012.
64. Nastase, S.A., Connolly, A.C., Oosterhof, N.N., Halchenko, Y.O., Guntupalli, J.S., Visconti di Oleggio Castello, M., Gors, J., Gobbini, M.I., and Haxby, J.V. (2017). Attention Selectively Reshapes the Geometry of Distributed Semantic Representation. Cereb. Cortex *27*, 4277–4291. https://doi.org/10.1093/cercor/bhx138.
65. Pitcher, D., Dilks, D.D., Saxe, R.R., Triantafyllou, C., and Kanwisher, N. (2011). Differential selectivity for dynamic versus static information in face-selective cortical regions. NeuroImage *56*, 2356–2363. https://doi.org/10.1016/j.neuroimage.2011.03.067.
66. Oosterhof, N.N., Tipper, S.P., and Downing, P.E. (2013). Crossmodal and action-specific: neuroimaging the human mirror neuron system. Trends Cogn. Sci. *17*, 311–318. https://doi.org/10.1016/j.tics.2013.04.012.
67. Centelles, L., Assaiante, C., Nazarian, B., Anton, J.L., and Schmitz, C. (2011). Recruitment of both the mirror and the mentalizing networks when observing social interactions depicted by point-lights: A neuroimaging study. PLoS One *6*, e15749. https://doi.org/10.1371/journal.pone.0015749.
68. Nikel, L., Sliwinska, M.W., Kucuk, E., Ungerleider, L.G., and Pitcher, D. (2022). Measuring the response to visually presented faces in the human lateral prefrontal cortex. Cereb. Cortex Commun. *3*, tgac036. https://doi.org/10.1093/texcom/tgac036.
69. Wang, Y., Metoki, A., Smith, D.V., Medaglia, J.D., Zang, Y., Benear, S., Popal, H., Lin, Y., and Olson, I.R. (2020). Multimodal mapping of the face connectome. Nat. Hum. Behav. *4*, 397–411. https://doi.org/10.1038/s41562-019-0811-3.
70. Julian, J.B., Fedorenko, E., Webster, J., and Kanwisher, N. (2012). An algorithmic method for functionally defining regions of interest in the ventral visual pathway. NeuroImage *60*, 2357–2364. https://doi.org/10.1016/j.neuroimage.2012.02.055.
71. Nunez-Elizalde, A., Deniz, F., la Tour, T.D., Castello, M.V.di O., and Gallant, J.L. (2021). Pymoten: motion energy features from video using a pyramid of spatio-temporal gabor filters. https://doi.org/10.5281/zenodo.4437446.
72. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). PyTorch: an imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. *32*, 8024–8035.
73. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.
74. Esteban, O., Markiewicz, C.J., Blair, R.W., Moodie, C.A., Isik, A.I., Erramuzpe, A., Kent, J.D., Goncalves, M., DuPre, E., Snyder, M., et al. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. Nat. Methods *16*, 111–116. https://doi.org/10.1038/s41592-018-0235-4.
75. Reuter, M., Rosas, H.D., and Fischl, B. (2010). Highly accurate inverse consistent registration: A robust approach. NeuroImage *53*, 1181–1196. https://doi.org/10.1016/j.neuroimage.2010.07.020.
76. Avants, B.B., Epstein, C.L., Grossman, M., and Gee, J.C. (2008). Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal. *12*, 26–41. https://doi.org/10.1016/j.media.2007.06.004.
77. Prince, J.S., Charest, I., Kurzawski, J.W., Pyles, J.A., Tarr, M.J., and Kay, K.N. (2022). Improving the accuracy of single-trial fMRI response estimates using GLMsingle. eLife *11*, e77599. https://doi.org/10.7554/eLife.77599.
78. Markiewicz, C.J., De La Vega, A., Wagner, A., Halchenko, Y.O., Finc, K., Ciric, R., Goncalves, M., Nielson, D.M., Kent, J.D., Lee, J.A., et al. (2022). poldracklab/fitlins: 0.11.0. https://doi.org/10.5281/zenodo.7217447.
79. Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., and Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. Front. Neuroinform. *8*, 14. https://doi.org/10.3389/fninf.2014.00014.
80. Lahner, B., Dwivedi, K., Iamshchinina, P., Graumann, M., Lascelles, A., Roig, G., Gifford, A.T., Pan, B., Jin, S., Murty, N.A.R., et al. (2023). BOLD Moments: modeling short visual events through a video fMRI dataset and metadata. https://doi.org/10.1101/2023.03.12.530887.
81. Groen, I.I., Greene, M.R., Baldassano, C., Fei-Fei, L., Beck, D.M., and Baker, C.I. (2018). Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. eLife *7*, e32962. https://doi.org/10.7554/eLife.32962.
82. Bellot, E., Abassi, E., and Papeo, L. (2021). Moving Toward versus Away from Another: how Body Motion Direction Changes the Representation of Bodies and Actions in the Visual Cortex. Cereb. Cortex *31*, 2670–2685. https://doi.org/10.1093/cercor/bhaa382.
83. Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. Annu. Rev. Vis. Sci. *1*, 417–446. https://doi.org/10.1146/annurev-vision-082114-035447.
84. Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848.
85. Arvai, K. (2020). Kneed. https://doi.org/10.5281/zenodo.6944485.
86. Yargholi, E., Hossein-Zadeh, G.A., and Vaziri-Pashkam, M. (2023). Two distinct networks containing position-tolerant representations of actions in the human brain. Cereb. Cortex *33*, 1462–1475. https://doi.org/10.1093/cercor/bhac149.
87. Dodell-Feder, D., Koster-Hale, J., Bedny, M., and Saxe, R. (2011). fMRI item analysis in a theory of mind task. NeuroImage *55*, 705–712. https://doi.org/10.1016/j.neuroimage.2010.12.040.
88. Brainard, D.H. (1997). The psychophysics toolbox. Spat. Vis. *10*, 433–436.
89. Pelli, D.G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spat. Vis. *10*, 437–442.

90. Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., and Broussard, C. (2007). What's new in psychtoolbox-3. Perception *36*, 1.

91. Esteban, O., Blair, R., Markiewicz, C.J., Berleant, S.L., Moodie, C., Ma, F., Isik, A.I., Erramuzpe, A., Kent, M., James, D., et al. (2018). fMRIPrep. Software. https://doi.org/10.5281/zenodo.852659.

92. Gorgolewski, K., Burns, C.D., Madison, C., Clark, D., Halchenko, Y.O., Waskom, M.L., and Ghosh, S.S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in Python. Front. Neuroinform. *5*, 13, https://doi.org/10.3389/fninf.2011.00013.

93. Gorgolewski, K.J., Esteban, O., Markiewicz, C.J., Ziegler, E., Ellis, D.G., Notter, M.P., Jarecka, D., Johnson, H., Burns, C., Manhães-Savio, A., et al. (2018). Nipype. Software. https://doi.org/10.5281/zenodo.596855.

94. Andersson, J.L.R., Skare, S., and Ashburner, J. (2003). How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. NeuroImage *20*, 870–888. https://doi.org/10.1016/S1053-8119(03)00336-7.

95. Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., and Gee, J.C. (2010). N4ITK: improved N3 bias correction. IEEE Trans. Med. Imaging *29*, 1310–1320. https://doi.org/10.1109/TMI.2010.2046908.

96. Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans. Med. Imaging *20*, 45–57. https://doi.org/10.1109/42.906424.

97. Dale, A.M., Fischl, B., and Sereno, M.I. (1999). Cortical surface-based analysis: I. Segmentation and surface reconstruction. NeuroImage *9*, 179–194. https://doi.org/10.1006/nimg.1998.0395.

98. Klein, A., Ghosh, S.S., Bao, F.S., Giard, J., Häme, Y., Stavsky, E., Lee, N., Rossa, B., Reuter, M., Chaibub Neto, E.C., et al. (2017). Mindboggling morphometry of human brains. PLOS Comput. Biol. *13*, e1005350, https://doi.org/10.1371/journal.pcbi.1005350.

99. Fonov, V., Evans, A., McKinstry, R., Almli, C., and Collins, D. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. NeuroImage *47*, S102. https://doi.org/10.1016/S1053-8119(09)70884-5.

100. Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. NeuroImage *17*, 825–841. https://doi.org/10.1016/s1053-8119(02)91132-8.

101. Cox, R.W., and Hyde, J.S. (1997). Software tools for analysis and visualization of fMRI data. NMR Biomed. *10*, 171–178. https://doi.org/10.1002/(SICI)1099-1492(199706/08)10:4/5<171::AID-NBM453>3.0.CO;2-L.

102. Greve, D.N., and Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. NeuroImage *48*, 63–72. https://doi.org/10.1016/j.neuroimage.2009.06.060.

103. Power, J.D., Mitra, A., Laumann, T.O., Snyder, A.Z., Schlaggar, B.L., and Petersen, S.E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. NeuroImage *84*, 320–341. https://doi.org/10.1016/j.neuroimage.2013.08.048.

104. Behzadi, Y., Restom, K., Liau, J., and Liu, T.T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. NeuroImage *37*, 90–101. https://doi.org/10.1016/j.neuroimage.2007.04.042.

105. Satterthwaite, T.D., Elliott, M.A., Gerraty, R.T., Ruparel, K., Loughead, J., Calkins, M.E., Eickhoff, S.B., Hakonarson, H., Gur, R.C., Gur, R.E., et al. (2013). An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. NeuroImage *64*, 240–256. https://doi.org/10.1016/j.neuroimage.2012.08.052.

106. Lanczos, C. (1964). Evaluation of noisy data. J. Soc. Ind. Appl. Math. Ser, B. Anal. *1*, 76–85. https://doi.org/10.1137/0701007.

107. Kay, K.N., Rokem, A., Winawer, J., Dougherty, R.F., and Wandell, B.A. (2013). GLMdenoise: a fast, automated technique for denoising task-based fMRI data. Front. Neurosci. *7*, 247.

108. Grossman, E., Donnelly, M., Price, R., Pickens, D., Morgan, V., Neighbor, G., and Blake, R. (2000). Brain areas involved in perception of biological motion. J. Cogn. Neurosci. *12*, 711–720. https://doi.org/10.1162/089892900562417.

109. Lescroart, M.D., Stansbury, D.E., and Gallant, J.L. (2015). Fourier power, subjective distance, and object categories all provide plausible models of BOLD responses in scene-selective visual areas. Front. Comput. Neurosci. *9*, 135. https://doi.org/10.3389/fncom.2015.00135.

Please cite this article in press as: McMahon et al., Hierarchical organization of social action features along the lateral visual pathway, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.10.015

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
raw fMRI data	This paper, OpenNeuro	https://doi.org/10.18112/openneuro.ds004542.v1.0.0
preprocessed fMRI, annotation, and eyetracking data	This paper, OSF	https://osf.io/4j29y/
Moments in Time dataset	Monfort et al.[18]	http://moments.csail.mit.edu
Probabilistic atlas of EVC and MT	Wang et al.[34]	https://napl.scholar.princeton.edu/document/66
Face, body, scene, and object parcels	Julian et al.[70]	https://web.mit.edu/bcs/nklab/GSS.shtml
Social perception regions	Deen et al.[47]	https://bendeen.com/data/
Software and algorithms		
custom code	This paper	https://doi.org/10.5281/zenodo.8381199
pymoten	Nunez-Elizalde et al.[71]	https://doi.org/10.5281/zenodo.6349625
Pre-trained AlexNet	PyTorch[72]	https://pytorch.org/hub/pytorch_vision_alexnet/
PyTorch	Paszke et al.[72]	https://pytorch.org
scikit-learn	Pedregosa et al.[73]	https://scikit-learn.org/stable/index.html
fMRIPrep	Esteban et al.[74]	https://doi.org/10.1038/s41592-018-0235-4
FreeSurfer	Reuter et al.[75]	https://freesurfer2016.sciencesconf.org
Advanced Normalization Tools	Avants et al.[76]	https://github.com/stnava/ANTsDoc
GLMsingle	Prince et al.[77]	https://doi.org/10.7554/eLife.77599
FitLins	Markiewicz et al.[78]	https://doi.org/10.5281/zenodo.7217447
Nilearn	Abraham et al.[79]	https://nilearn.github.io/stable/index.html

RESOURCE AVAILABILITY

Lead contact

Further information or access to the stimuli in the current study, please contact the corresponding author, Emalie McMahon, emaliemcmahon@jhu.edu.

Materials availability

The only material contribution of this study is the video stimulus set. Because of MiT licenses restrictions, please email the corresponding author for access to the videos used in the current study.

Data and code availability

We have made all data and code for this project available for others to use. See key resources table for links.

Please cite this article in press as: McMahon et al., Hierarchical organization of social action features along the lateral visual pathway, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.10.015

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Video annotation participants

Annotations were collected on the Prolific platform. Only American participants with normal or correct-to-normal vision and with a prior approval rate greater than 85% were eligible to join. Prior to participation, participants (n = 2279) gave informed consent. The Johns Hopkins University Press Institutional Review Board approved the consent and protocol. Participants received monetary compensation for their time (rate of $10 per hour). Demographic data was saved from Prolific at a later date than responses on the main task resulting in some missing demographic data; subject numbers report the number of participants with demographic data [Gender (n = 2014): female = 979, male = 1121, prefer not to say = 4; Age (n = 2085): M = 36.26, SD = 12.44)] .

fMRI participants

fMRI data were collected in four participants (Females = 2, mean age 25.5 years, range 23–30 years, 3 Caucasian and 1 Asian). All participants were right-handed and had normal or corrected-to-normal vision. Participants gave written informed consent before participation and were monetarily compensated for their time. The Johns Hopkins School of Medicine Institutional Review Board approved the consent and protocol.

Eye tracking participants

Thirteen participants participated in the eye tracking experiment (9 females, 4 males, mean age 26 years, range 18-35 years, 7 Caucasian, 5 Asian, and 1 other). Two participants were not analyzed due to poor data quality during the session: one participant fell asleep, and the other's pupil was not reliably identified by the eye tracker throughout the experiment. Participants had normal vision or sufficient visual acuity to appreciate the videos at the presented distance. Participants gave written informed consent before participation and were monetarily compensated for their time. The Johns Hopkins University Institutional Review Board approved the consent and protocol.

METHOD DETAILS

Stimulus set

Stimulus selection

The stimulus set is a subset of the MiT dataset.[18] MiT is a large dataset containing many social action categories. For this reason, previous cognitive neuroscience work has also used subsets of MiT to study event[80] and social action perception.[1] We procedurally removed videos to select a set reasonable for a cognitive neuroscience experiment (n = 250). We first removed action categories that were unlikely to contain human actions (e.g., feeding or bubbling) or human actions that were not common-place actions as defined by the American Time Use Survey (e.g., spitting).[19]

Following this, we reduced the stimulus set to videos with only two people (a critical distinction from the abovementioned cognitive neuroscience studies using MiT) because the number of people is a reliable indicator of sociality.[1,22] Specifically, we used Amazon's Rekognition face-detection algorithm to select the videos that the model determined to have exactly two people with 95% confidence. From this subset, we manually removed videos with greater than or fewer than two people leaving 2,751 videos. To make social interpretations straightforward, we then removed videos with staged actions (e.g., instructional or stock videos) and videos showing a person speaking to someone off-camera. We also removed videos based on these criteria: animation, obvious scene cuts, watermarks/logos, obvious temporal distortions (either slow-motion or time-lapse), and low spatial or temporal resolution.

This left 723 videos. We resampled videos to be 30 Hz and have precisely ninety frames. We also center-cropped videos to be square and resized all videos to 500 x 500 pixels. Center-cropping removed a person from the video in some cases so these videos were removed (n = 3).

On this 720 video dataset, we collected annotations of several dimensions (spatial expanse, object directedness, agent distance, communication, joint action, intimacy, dominance, cooperation, valence and arousal). These dimensions were based off prior neuroimaging of social actions[1–3,22] and social features that may be extracted visually.[29] Three dimensions (intimacy, dominance, and cooperation) were ultimately not included in the final analyses because preliminary encoding models using cross-validation within the fMRI training set did not reveal significant prediction in ROIs across subjects.

Based on previous literature, we anticipated that joint action would be the primary feature of social interactions represented in the brain,[28] but we wanted detection of joint action to not be trivial based on low- or mid-level visual cues. As a result, we chose to select the videos for the final stimulus set that would reduce the rank correlation between joint action and agent distance ($rho(718) = -0.4$) as has been done in previous work.[81] Separately for indoor and outdoor videos, we randomly removed videos ($n_{indoor} = 120$, $n_{outdoor} = 100$) with agent distance rating below the mean and joint action rating above the mean. The resulting rank correlation was reduced ($rho(498) = -0.3$).

From this set (n = 500), the final stimulus set shown in the scanner (n = 250) was selected to remove videos with camera panning or excessive camera motion. The training-test split was performed by randomly splitting the dataset in two hundred and 50 video sets. To ensure that the test set was not out of distribution of the training set, the distribution of ratings for key features (indoor, spatial expanse, and joint action) was visually inspected to determine whether the spread of these feature values was qualitatively similar in the training and test sets. This process was repeated until a good split was found.

Please cite this article in press as: McMahon et al., Hierarchical organization of social action features along the lateral visual pathway, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.10.015

Annotations of the facingness feature were collected after the selection of the final set based on emerging evidence that facingness is represented in EBA.[10,11,82]

Annotations

Participants rated 40 videos on a single feature on a Likert Scale from one to five. This was done for each of the features in Figure 1 with two exceptions: facingness, which was collected later than the other features, was rated in groups of 25 videos, and indoor was rated only by the first author. Videos appeared one at a time at a resolution of 500 x 500 pixels at a frame rate of 30 Hz. They had unlimited time to respond but had to respond to continue with the experiment.

We removed participants with incomplete data (n = 225) or who used only one or two of the five Likert options across all videos (n = 41). Additionally, we iteratively excluded participants if the correlation of their responses with that of other participants was more than three standard deviations away from the mean (n = 62).

Following data cleaning, ratings were min-max scaled to be in the range of zero-to-one from the Likert range of one-to-five. The average rating for each video was treated as a single-dimensional representation of a particular feature for that video and used in the subsequent encoding models.

Algorithmic feature extractions

Because AlexNet-conv2 is a good model of EVC,[83] low-level visual features were estimated using the second convolutional (AlexNet-conv2) layer of an ImageNet[84] trained AlexNet.[20] We extracted activations from AlexNet-conv2 using PyTorch[72] for each frame and then averaged activations across all frames of the three-second video. The dimensionality of the features was reduced using principal components analysis (PCA) learned in the training set and applied to the test set. The number of PCs needed to reach the elbow was calculated algorithmically with *knee*.[85]

Motion energy was estimated with an Adelson and Bergen model[21] implemented in *pymoten*[71] using the default pyramid with a temporal window of 10 frames. The motion energy was then averaged across spatiotemporal windows. PCA was again used to reduce dimensionality (PCs = 3).

fMRI Experiment

Multi-session scanning

Scanning for each participant took place over four separate 2 h sessions. During these sessions, three participants completed 60 runs of the experiment (10 repeats of training videos, 20 of test videos), and one completed 54 runs (9 repeats of training videos, 18 of test videos). High-resolution anatomical images were collected during each session for EPI registration. Participants completed a battery of functional localizer tasks during the first scanning session.

Faces, bodies, objects, and scenes functional localizer

To localize face, body, and object regions, participants completed three runs of a dynamic localizer from Pitcher et al.[65] Three of the four participants saw blocks with faces, bodies, objects, scenes, and scrambled objects (duration = 414 s). One participant only saw faces, bodies, objects, and scenes (duration = 342 s).

Social interaction functional localizer

Participants completed three runs of a social interaction localizer using point light figures either interacting (social interactions) or performing independent actions (nonsocial actions) from Isik et al.[4] The duration of each scan was 159 s.

Biological motion functional localizer

Participants completed two runs of a biological motion localizer from Yargholi et al.[86] The duration of the scans was 435 s. The task was composed of intact point-light figures, position-scrambled point-light figures, random translational motion, and static point-light figures.

Theory of mind functional localizer

To localize theory of mind regions, participants completed two runs of the false belief task based on Dodell-Feder et al.[87] Each run lasted for 273 s.

Main task procedure

Participants in the scanner viewed 250 three-second videos. Following annotation, we separated the videos into training (two hundred videos) and test sets (fifty videos) such that the distribution of ratings for joint action and indoor was similar between training and test sets.

The experiment was designed in sections of six fMRI runs. Within a section, the two hundred training videos were randomly separated into four runs of fifty videos. To generate a highly reliable test set for model estimation, we presented test videos in separate runs twice per section.

Within a run, the order of videos was always randomized. Participants freely viewed fifty dyadic videos and five randomly interspersed ''crowd'' videos containing many people. To ensure participants remained attentive, they hit a button for every crowd presentation. Every participant detected the crowd videos with greater than 99% accuracy. Following beta-weight estimation in the GLM, we excluded crowd videos from all subsequent analyses.

The videos were each shown for 3 s at a resolution of 500 x 500 pixels and a frame rate of 30 Hz. A black screen was shown between each video for 1.5 s. On a random ten trials per run, an additional 1 TR (1.5 s) jitter was added to the ISI timing. The first stimulus was presented 2 TRs after the start of the first steady-state volume, and the run ended 9 TRs after the final video. The total run time was 279 s.

Please cite this article in press as: McMahon et al., Hierarchical organization of social action features along the lateral visual pathway, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.10.015

fMRI acquisition parameters

All fMRI scans were conducted on a 3T Philips Elition RX scanner (with a 32-channel head-coil) at the F.M. Kirby Research Center for Functional Brain Imaging at the Kennedy Krieger Institute.

During the first scanning session, we collected two anatomical scans. One was collected with an axial primary slice direction and the other with a sagittal primary slice direction. The sagittal scan only occurred during the first session. The axial scan was repeated during every session. All five anatomical scans were combined with *fMRIPrep*.

The axial scans were performed with a T1-weighted magnetization-prepared rapid-acquisition gradient-echo sequence with the following parameters: repetition time (TR) = 8.0 ms, echo time (TE) = 3.7 ms, flip angle = 8°, voxel size = 0.95 x 0.95 x 1 mm^3 field of view = 224 x 224 x 150 mm^3.

The sagittal scans were performed with a T1-weighted magnetization-prepared rapid-acquisition gradient-echo sequence with the following parameters: repetition time (TR) = 7.6 ms, echo time (TE) = 2.4 ms, flip angle = 18°, voxel size = 1 x 1 x 1 mm^3 field of view = 240 x 240 x 180 mm^3

T2*-weighted functional data were acquired using a multi-band (factor 4) gradient-echo echo-planar imaging sequence with the following parameters: repetition time (TR) = 1.5 s, echo time (TE) = 30 ms, flip angle = 52°, voxel size = 2 x 2 x 1.93 mm^3, field of view = 216 x 120 mm^2, and 60 axial slices spanning across the entire cortex.

Eye tracking experiment

Apparatus

Videos were presented using Psychtoolbox,[88–90] while eye tracking data were collected with a tower-mounted EyeLink CL 4.56 eye-tracker (SR Research Ltd, 2012). Videos were displayed on a 435 mm x 240 mm monitor (1600 x 900 pixels) in the center of the screen at a resolution of 900 x 900 pixels and extended 21 degrees of visual angle. Participants' monocular (right eye) gaze was tracked remotely at a sampling rate of 500 Hz, while they were seated in a chin rest.

Procedure

The experiment consisted of 440 trials overall consisting of the 50 videos in the test set (presented 8 times each with repetitions across blocks) and 40 catch trials. Before the experiment, participants were told of the structure of the experiment, to hit the any button for the catch trials, and otherwise to watch the videos normally. The experiment was divided into halves with 9-point calibration performed before each half. The experiment took place in a darkened room with the experimenter in a separate room monitoring the eye tracking quality and controlling calibration.

One block of the experiment consisted of viewing all 50 videos in a random order. On an additional 5 trials during the block, participants saw videos with a crowd of people. As in the fMRI experiment, participants were instructed to push a button on these trials. Videos were presented for their duration (3 s) with 750 ms of blank screen between videos. Between blocks, participants had self-timed breaks during which they were instructed to rest their eyes but stay in the chin rest. At the half-way point, participants were allowed to sit back and tell the experimenter when they were ready to continue.

QUANTIFICATION AND STATISTICAL ANALYSIS

Feature regression

To quantify the relation among all features of our stimulus set, we performed pairwise regression between all pairs of features. We learned a linear mapping between the predictor and predicted feature in the training set and predicted the response in the test set. The sign-squared correlation between the predicted response and true rating defines the strength of the relation between each pair of features. We used permutation testing to test significance by randomly shuffling stimulus labels in the test set and repeating the prediction procedure five thousand times to estimate a null distribution. To estimate variance, we performed bootstrapping over random paired samples of the data over five thousand resamples.

Because AlexNet-conv2 and motion energy are multidimensional while all other features are one dimensional, they were always used as the predictor of the annotated features. Further, to establish the relation between AlexNet-conv2 and motion energy, the motion energy PCs were averaged prior to prediction.

fMRI Preprocessing

fMRIPrep

Results included in this manuscript come from preprocessing performed using fMRIPrep 21.0.2,[74,91] which is based on ***Nipype*** 1.6.1.[92,93]

The next three sections (B_0 inhomogeneity mappings, anatomical data, and functional data) of boilerplate text were automatically generated by fMRIPrep with the express instructions that users should copy and paste this text into their manuscripts *unchanged*. It is released under the CC0 license.

B_0 inhomogeneity mappings

A total of 4 fieldmaps were found available within the input BIDS structure for this particular subject. A ***B_0***-nonuniformity map (or ***FieldMap***) was estimated based on two (or more) echo-planar imaging (EPI) references with topup.[94]

Please cite this article in press as: McMahon et al., Hierarchical organization of social action features along the lateral visual pathway, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.10.015

Anatomical data

A total of 5 T1-weighted (T1w) images were found within the input BIDS dataset. All of them were corrected for intensity non-uniformity (INU) with N4BiasFieldCorrection,[95] distributed with ANTs 2.3.3.[76] The T1w-reference was then skull-stripped with a ***Nipype*** implementation of the antsBrainExtraction.sh workflow (from ANTs), using OASIS30ANTs as target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast (FSL 6.0.5.1:57b01774[96]). A T1w-reference map was computed after registration of 5 T1w images (after INU-correction) using mri_robust_template (FreeSurfer 6.0.1[75]). Brain surfaces were reconstructed using recon-all (FreeSurfer 6.0.1[97]), and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle.[98] Volume-based spatial normalization to one standard space (MNI152NLin2009cAsym) was performed through nonlinear registration with antsRegistration (ANTs 2.3.3), using brain-extracted versions of both T1w reference and the T1w template. The following template was selected for spatial normalization: ***ICBM 152 Nonlinear Asymmetrical template version 2009c***[99] (TemplateFlow ID: MNI152NLin2009cAsym).

Functional data

For each of the 72 BOLD runs found per subject (across all tasks and sessions), the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated by aligning and averaging 1 single-band references (SBRefs). Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using mcflirt (FSL 6.0.5.1:57b01774[100]). The estimated ***FieldMap*** was then aligned with rigid-registration to the target EPI (echo-planar imaging) reference run. The field coefficients were mapped on to the reference EPI using the transform. BOLD runs were slice-time corrected to 0.7s (0.5 of slice acquisition range 0s-1.4s) using 3dTshift from AFNI.[101] The BOLD reference was then co-registered to the T1w reference using bbregister (FreeSurfer) which implements boundary-based registration.[102] Co-registration was configured with twelve degrees of freedom to account for distortions remaining in the BOLD reference. First, a reference volume and its skull-stripped version were generated using a custom methodology of ***fMRIPrep***. Several confounding time-series were calculated based on the ***preprocessed BOLD***: framewise displacement (FD), DVARS and three region-wise global signals. FD was computed using two formulations following Power et al.[103] (absolute sum of relative motions) and Jenkinson et al.[100] (relative root mean square displacement between affines) FD and DVARS are calculated for each functional run, both using their implementations in ***Nipype*** (following the definitions by Power et al.[103]). The three global signals are extracted within the CSF, the WM, and the whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction (***CompCor***[104]). Principal components are estimated after high-pass filtering the ***preprocessed BOLD*** time-series (using a discrete cosine filter with 128s cut-off) for the two ***CompCor*** variants: temporal (tCompCor) and anatomical (aCompCor). tCompCor components are then calculated from the top 2% variable voxels within the brain mask. For aCompCor, three probabilistic masks (CSF, WM and combined CSF+-WM) are generated in anatomical space. The implementation differs from that of Behzadi et al. in that instead of eroding the masks by 2 pixels on BOLD space, the aCompCor masks are subtracted a mask of pixels that likely contain a volume fraction of GM. This mask is obtained by dilating a GM mask extracted from the FreeSurfer's ***aseg*** segmentation, and it ensures components are not extracted from voxels containing a minimal fraction of GM. Finally, these masks are resampled into BOLD space and binarized by thresholding at 0.99 (as in the original implementation). Components are also calculated separately within the WM and CSF masks. For each CompCor decomposition, the ***k*** components with the largest singular values are retained, such that the retained components' time series are sufficient to explain 50 percent of variance across the nuisance mask (CSF, WM, combined, or temporal). The remaining components are dropped from consideration. The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file. The confound time series derived from head motion estimates and global signals were expanded with the inclusion of temporal derivatives and quadratic terms for each.[105] Frames that exceeded a threshold of 0.5 mm FD or 1.5 standardized DVARS were annotated as motion outliers. The BOLD time-series were resampled into standard space, generating a ***preprocessed BOLD run in MNI152NLin2009cAsym space***. First, a reference volume and its skull-stripped version were generated using a custom methodology of ***fMRIPrep***. The BOLD time-series were resampled onto the following surfaces (FreeSurfer reconstruction nomenclature): ***fsnative***. All resamplings can be performed with ***a single interpolation step*** by composing all the pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and output spaces). Gridded (volumetric) resamplings were performed using antsApplyTransforms (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels.[106] Non-gridded (surface) resamplings were performed using mri_vol2surf (FreeSurfer).

Many internal operations of ***fMRIPrep*** use ***Nilearn*** 0.8.1,[79] mostly within the functional processing workflow. For more details of the pipeline, see the section corresponding to workflows in fMRIPrep's documentation.

fMRI GLM

Functional localizer GLM

FitLins[78] was used to run the general linear model (GLM) and compute first-level analyses for each localizer task. Block presentations were convolved with a Statistical Parametric Mapping (SPM) hemodynamic response function (HRF). Translation and rotation motion parameters were included. Data were smoothed using a 4 mm FWHM kernel.

Please cite this article in press as: McMahon et al., Hierarchical organization of social action features along the lateral visual pathway, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.10.015

Main task GLM

Following preprocessing, data were smoothed using a 3 mm FWHM kernel implemented in the *smooth_img* function from nilearn.[79] Although the EPI data were morphed to a common space by default in fMRIPrep, all analyses took place in the subject's native volumetric space and were morphed to an individual subject's surface only for visualization.

Because of the denoising procedure of *GLMsingle*, we did not include motion correction parameters in the GLM. *GLMsingle*[77] was used to estimate the individual trial responses for the main fMRI experiment. We used all three processing stages of *GLMsingle*: fitting a library of HRFs in individual voxels, *GLMdenoise*,[107] and fractional ridge regression. Due to memory considerations, we fit the GLM separately for each participant and session. To prevent contamination of the training set by the test set, we did not allow *GLMsingle* to scale the final estimates or convert the values into percent signal change. Instead, within a particular session, the training data were normalized, and the test data were normed by the learned mean and standard deviation from training.

ROI definition in native space

For each region, parcels from an appropriate source (see below) were first warped from MNI space into each subject's native space using *antsApplyTransforms*.[76] Following the definition of each ROI, we removed overlapping voxels between ROIS as well as voxels outside of the reliability mask. We combined ROIs across hemispheres.

Though we included theory of mind and biological motion tasks, we do not include ROI results from these localizers. In the case of theory of mind, while we were able to consistently localize the temporal parietal junction (TPJ) in individual subjects, we found in early analyses that the localized TPJ fell almost entirely outside the reliability mask, which is also true of other studies of naturalistic action recognition.[3] Because our ROI analysis was limited to voxels within the ROI and reliability mask, we chose not to model responses in the TPJ. For the biological motion task on the other hand, we were not able to localize the STS biomotion region in individual subjects using either a biological motion minus scrambled biological motion[108] or minus translation motion contrast.[86] Others have successfully localized biomotion-STS using biological motion minus rigid body motion,[47] but we did not include the rigid body moiton condition in our localizer. Due to our inability to find consistent biomotion responses, we do not report biomotion-STS results here.

Anatomical ROIs (EVC and MT)

Probabilistic parcels from Wang et al.[34] were first warped from MNI space into each subject's native space using *antsApplyTransforms* from ANTS.[76] EVC was defined as the combination of the V1v, V1d, V2v, and V2d probabilistic parcel. MT was defined as the full probabilistic parcel. The Wang et al.[34] atlas is defined in fsaverage space. Due to variability in sulcal folding in individuals in this region of cortex, this method of defining MT using the atlas is a limitation in the current study.

Anatomically constrained functional ROIS (pSTS-SI and aSTS-SI)

Looking at the contrast between social interactions and nonsocial actions in the point light figures from Isik et al.,[4] we noticed that the most active voxels were in more anterior regions than previously reported.[4] For this reason, we defined two separate social interaction (SI) functional ROIs by dividing the STS parcel from Deen et al.[47] in half along the posterior-anterior axis. We defined both posterior (pSTS-SI) and anterior (aSTS-SI) as the top ten percent most active voxels from the social interaction minus nonsocial action contrast.

Functional ROIs (FFA, STS-Face, PPA, EBA, and LOC)

We used ROI parcels from Julian et al.[70] and Deen et al.[47] to localize common visual category selective regions. We removed voxels that were not present in an ROI in at least half of the original subjects. The top ten percent of voxels within the parcel was defined as the ROI in an individual participant. From the face, body, object, scene task, contrasts were defined to localize FFA and STS-Face (faces minus objects), PPA (scenes minus objects), EBA (bodies minus objects), and LOC (objects minus scrambled objects). For sub-01, there were no scrambled objects presented in the localizer, so LOC was defined using an objects-minus-scenes contrast.

Voxel-wise encoding models

Split-half reliability

The quality of the data in the test set was evaluated by computing the split-half reliability within-subject. We averaged response from interleaved runs and calculated the correlation between the two halves of the data. The reliability of the data was liberally thresholded at the critical value for a one-tailed uncorrected significant correlation ($r(48) = 0.117$, $p < 0.05$).

Model training and evaluation

Both model training and evaluation were done within-subject. We used cross-validation in the training set to finalize the modeling procedure including decisions such as fMRI preprocessing decisions and using OLS regression and variance partitioning. Analyses reported here on the held-out test set.

Before model fitting, we masked neural data to the reliable voxels and averaged across repetitions. We normalized features for the training videos, and the test set was normed by the mean and standard deviation of the training set.

Because the maximum number of features (20 AlexNet-conv2, 3 motion energy, and 9 annotated features) relative to the number of samples (200 videos) was low, we fit an ordinary least squares (OLS) model. Others have also argued that regularized regression, though more common in voxel-wise encoding analyses, complicates variance partitioning analyses.[109] Thus, we fit the OLS model in training data and predicted the response in the test data using the *LinearRegression* from scikit-learn.[73] The sign-squared correlation between the predicted response and true response was the measure of model performance. We opted for the sign-squared correlation as opposed to simply the correlation in order to maintain consistency between standard and variance partitioning analyses as the latter requires squaring. However, squaring the correlation without maintaining the sign result in the same numerical value

Please cite this article in press as: McMahon et al., Hierarchical organization of social action features along the lateral visual pathway, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.10.015

for both good prediction ($r = 0.5$, $r^2 = 0.25$, signed $r^2 = 0.25$) and extremely poor prediction ($r = -0.5$, $r^2 = 0.25$, signed $r^2 = -0.25$), which is particularly problematic in computing the permuted null distribution.

We used permutation testing to test significance by randomly shuffling stimulus labels in the test set and repeating the prediction procedure ten thousand times to estimate a null distribution. To estimate variance, we performed bootstrapping over random paired samples of the data over ten thousand resamples.

We fit several models described in detail in the corresponding Results sections.

ROI analysis

Following voxel-wise encoding (for all models used), we averaged the prediction across each ROI in each subject. We did the same for the distributions estimated by permutation testing and bootstrapping. To determine the significance at the ROI level, we compared the observed ROI-average prediction to the estimated ROI-average null distribution and calculated one-tailed probability of the observed result in the null distribution. Within subject and ROI, we FDR corrected for multiple comparisons across categories of features or individual features.

To calculate the group-level ROI results, we calculated the average prediction and distributions estimated by permutation testing and bootstrapping across subject within each ROI. To calculate significance, we computed the one-tailed probability of the observed average response across subjects given the estimated null distribution across subjects. Within an ROI, we used FDR correction to correct for multiple comparisons across features or category of features.

Variance partitioning

To determine whether a given feature category or individual feature explained unique variance in a given ROI, we used variance partitioning. We did this by taking the variance explained by the full model with all features minus the variance explained by the full model without the category or feature of interest. We again used permutation testing and bootstrapping (ten thousand iterations each) to estimate the significance and variance of the model performance, respectively.

Whole-brain preference maps and feature prediction

Voxel-wise significance was computed as described in model training and evaluation. We corrected the p-values for multiple comparisons using FDR correction. Whole-brain results are the visualization of these results morphed to the cortical surface.

The preference maps were computed by first finding the voxels that were significantly predicted by a single category and labeled according to that category. Then if any voxel was predicted by more than one category, we simply found which category maximally predicted the voxels and assigned that label to the voxel. We then morphed this to the surface of each individual subject and visualized the result.

Evaluating face-related effects

Nonparametric ANOVA

To compare the relative responses between categories and ROIs, we computed a nonparametric ANOVA. To do this we calculated the difference in each of two ROIs between the prediction accuracies of two categories of interest, and then calculated the difference of differences between ROIs. To assess significance, we performed permutation testing by shuffling video labels and repeating the above procedure.

Face-feature correlation

In a post-hoc analysis, we asked directly whether descriptors of the face size and location correlated with our annotated features. We computed the correlation between each feature and face area (the sum of the area of the two face bounding boxes averaged across frames) and centrality (the minimum distance of the two bounding boxes from the center of the frame averaged across frames). We then tested for significance by computing the permuted null distribution over five thousand iterations. Finally, we corrected for multiple comparisons across features using FDR correction.

Eye tracking

Heatmaps

To summarize the pattern of fixations, we computed a heatmap of participant's fixation on each trial with no more than 30% of samples missing (as the result of blinks or lost tracks), which corresponds to around 1 s of the 3s video. This resulted in the removal of a small number of trials on average across participants ($M = 5.30\%$, $SD = 3.93\%$). The heatmap was computed as the 2D histogram of the eye-tracking samples with twenty bins along each dimension (400 total bins).

Within- and between-subject reliability

To compute the within-subject reliability, we averaged the heatmaps on every other presentation and computed the Pearson correlation between the heatmaps on odd and even repetitions for every video. We report the within-subject reliability averaged across all videos.

For the between-subject reliability, we averaged the heatmaps across all presentations of each video. Then for a given participant, we correlated their heatmap with the average of all other subjects' heatmap for a given video and repeated for every video. We report the between-subject reliability averaged across videos.

Proportion of time looking at faces

To compute the proportion of time that participants looked faces, we summed the heatmap cells that overlapped with the face bounding boxes for each video and divided by the total number of usable samples. We averaged the proportions across repetitions of the same video.

Please cite this article in press as: McMahon et al., Hierarchical organization of social action features along the lateral visual pathway, Current Biology (2023), https://doi.org/10.1016/j.cub.2023.10.015

Feature correlation

We related the feature annotations to the eye tracking data in two ways. First, we asked whether the consistency in viewing pattern across repetitions of the same video was related to the content of the video. Second, we asked whether the proportion of time spent looking at faces was related to the content of the video. To address both of the questions, we computed the Pearson correlation between the feature ratings and the eye tracking data (either the within-subject correlation or the proportion looking at faces). We used permutation testing to test significance within-subject by randomly shuffling stimulus labels in the test set and repeating the prediction procedure five thousand times to estimate a null distribution. To calculate group-level significance, we computed the two-tailed probability of the observed average response across subjects given the estimated null distribution across subjects. We corrected for multiple comparisons across features using FDR correction.

Spatial expanse and communication reliability comparison

To compare whether within-subject reliability was more related to spatial expanse or communication. We first computed the absolute value of the correlation of reliability with spatial expansion and communication (and the corresponding permuted distributions). While it is meaningful that the correlation between spatial expanse is negative (participants looking pattern becomes less reliable as the spatial expanse becomes larger), here we are only interested in whether the magnitude of the relation is greater for spatial expanse than communication. Thus, we compared the difference in magnitude to the difference in the null distribution and computed the two-tailed probability of the difference.