

# Neural signatures of emotional intent and inference align during social consensus

Received: 10 May 2024

Accepted: 6 May 2025

Published online: 08 July 2025

 Check for updates

Marianne C. Reddan  , Desmond C. Ong<sup>3</sup>, Tor D. Wager  ,  
Sonny Mattek , Isabella Kahhale<sup>5</sup> & Jamil Zaki<sup>2</sup> 

Humans effortlessly transform dynamic social signals into inferences about other people's internal states. Here we investigate the neural basis of this process by collecting fMRI data from 100 participants as they rate the emotional intensity of people (targets) describing significant life events. Targets provide self-ratings on the same scale. We then train and validate two unique multivariate models of observer brain activity. The first predicts the target's self-ratings (i.e., intent), and the second predicts observer inferences. Correspondence between the intent and inference models' predictions on novel test data increases when observers are more empathically accurate. However, even when observers make inaccurate inferences, the target's intent can still be predicted from observer brain activity. These findings suggest that an observer's brain contains latent representations of other people's socioemotional intensity, and that fMRI models of intent and inference can be combined to predict empathic accuracy.

Healthy social functioning increases life expectancy<sup>1</sup>, buffers cognitive decline<sup>2,3</sup>, improves mental health conditions like depression<sup>4</sup>, and enriches a person's daily life<sup>5</sup>. The quality of a social interaction is shaped by two essential behaviors—how we signal our emotions and how we infer the emotions of others<sup>6</sup>. People who make clear signals and accurate inferences tend to have healthy adolescent adjustment, stable relationships, and high subjective well-being<sup>7,8</sup>. Conversely, people who make ambiguous signals and inaccurate inferences are more likely to experience social isolation, and, in severe cases, may meet criteria for clinical conditions like autism spectrum disorder<sup>9</sup>, schizophrenia<sup>10</sup>, social anxiety<sup>11</sup>, and borderline personality disorder<sup>12</sup>. Understanding how the brain processes socioemotional signals and generates inferences can help pinpoint whether errors arise during signal perception or inference-making. This understanding could aid the development of personalized interventions that reduce loneliness and improve public health<sup>13</sup>.

The neural processes that support conscious inference are computationally complex and require the integration of multiple sources of information such as an observer's internal homeostatic

state, memories of their past experiences, expectations, and social schemas<sup>14,15</sup>. Neuroimaging studies have implicated the amygdala, medial prefrontal cortex (mPFC), temporoparietal junction (TPJ), and precuneus<sup>16–18</sup> in socioemotional inference; however, models of how multiple brain regions interact with one another to form an inference are lacking. Prior research also fails to disentangle signal intent (the signaler's intended message) from inference (the observer's interpretation). Instead of intent, most research uses abstracted measures such as normative ratings and consensus judgments<sup>19–21</sup>. This is partly because self-reported intent is not available when stimuli are derived from normative data sets or actors. The importance of studying intent is subtle, but significant. Self-reported intent ratings are “intimate” to the signaler's internal processes and the meaning they wish to convey. The goal of natural socioemotional signaling is for this intimate meaning to be understood by the observer; not for the observer to match their inferences with social norms<sup>6</sup>. Mutual understanding is the foundation of a healthy relationship and promotes interpersonal cooperation<sup>13</sup>. The Stanford Emotional Narratives Dataset (SENDv1)<sup>22</sup> provides

<sup>1</sup>Department of Psychiatry and Behavioral Sciences, Albert Einstein College of Medicine, Bronx, NY, USA. <sup>2</sup>Department of Psychology, Stanford University, Stanford, CA, USA. <sup>3</sup>Department of Psychology, University of Texas at Austin, Austin, TX, USA. <sup>4</sup>Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, USA. <sup>5</sup>Department of Psychology, University of Pittsburgh, Pittsburgh, PA, USA.  e-mail: [marianne.reddan@einsteinmed.edu](mailto:marianne.reddan@einsteinmed.edu); [jzaki@stanford.edu](mailto:jzaki@stanford.edu)

self-reported intent ratings, but has not been used in an fMRI study until now.

Previous work correlating observer inferences to targets' self-reports found that accuracy is associated with higher activation in the inferior parietal lobule (IPL), premotor cortex, superior temporal sulcus (STS), and mPFC<sup>18</sup>. Accuracy, however, is not a neural process, but an outcome of social signal processing and inference formation. That is, accuracy is determined by how the signaler compresses the information they intend to convey into a recognizable social signal (e.g., a grimace or lowering of the eyes), and how that signal is perceived and interpreted by an observer. To better understand the neural mechanisms supporting accuracy in this task, it is critical to disentangle the signaler's intent from the observer's inference and to model the distributed patterns of brain activity underlying each component. Such models can then be combined to explain the factors that drive individual differences in inference accuracy.

Naturalistic stimuli are necessary for the study of social signal processing<sup>23</sup>. Though static and unimodal stimuli are useful for isolating specific components of emotion perception and induction, low-dimensional stimuli forgo the complexity of the dynamic environments we live in and limit our ability to study social behavior<sup>23–25</sup>. Naturalistic audiovisual recordings allow us to capture the subtleties of nonverbal communication, such as nuanced facial expressions, tone of voice, and body language. Moreover, naturalistic stimuli reduce demand characteristics<sup>26</sup> and allow us to examine the context in which social interactions occur. For instance, studying how people recognize distress in real-life situations provides a better understanding of empathic inference than asking people to rate emotions in a hypothetical situation. First-person narratives, such as storytelling, are particularly opportune stimuli for the study of real-world social interactions<sup>23</sup>. Such stimuli can be sufficiently constrained in the laboratory by limiting spurious background activity while still allowing for the natural unfolding of dynamic socioemotional signaling<sup>22</sup>. Overall, studying naturalistic stimuli can lead to a more accurate understanding of real-world social interactions, which can inform the development of treatments for social and emotional disorders.

To improve our understanding of how an observer's brain perceives signals and forms inferences in naturalistic contexts, here we collect videos of people (targets) describing emotional events in their lives (see Figs. S1 & S2). Immediately after recording themselves, targets provide moment-by-moment ratings of how they felt as they spoke. These self-reports constitute the "intent" of their social signals. We then scan a second set of participants (observers) with fMRI while they view targets and rate what they thought the target was feeling at each moment on the same scale (see Fig. 1a; Figs. S3–8). These ratings constitute the observer's live inferences. Correspondence between observer inference and target self-report (intent) serve as a measure of empathic accuracy, or agreement between two people about what one of them feels. Both sets of ratings are normalized and constrained to five continuous levels of socioemotional intensity and used in a regression model to predict changes in participant-level fMRI data (see Method detail). The resulting coefficient maps are then fed to a regression-based machine learning algorithm to develop two neural signatures of socioemotional processing— one indicative of the signaler's intent (i.e., the target's self-report) and one indicative of the observer's inference. The models are cross-validated during training and externally validated on held-out data after training. Finally, we test if the models can be used concurrently to predict an observer's empathic accuracy and explore brain regions involved in the transition from signal perception to a reportable inference.

Here we show, through high-quality naturalistic stimuli that reflect real-world experiences, multivariate methods, and the self-reported intentions of social targets, that two unique components of real-world socioemotional inference can be predicted from observer brain activity and combined to explain an individual's empathic accuracy.

## Results

### Socioemotional intent can be predicted from observer brain activity

We developed a whole-brain fMRI model that can predict a target's self-reported emotional intensity (i.e., signal intent) from an observer's brain activity. The model's features (X) included the observers' ( $N=100$ ) signal intent whole-brain coefficient maps, which were derived for each intensity quintile via participant-level general linear models (GLMs) of each observer's timeseries fMRI data (Fig. 1a–c; see Methods). Each observer had a set of five coefficient maps. The dependent variable (Y) was the quintile's numeric label (1–5). The model was trained using least absolute shrinkage and selection operator-regularized principal components regression (LASSO-PCR)<sup>27</sup> and leave-one-subject-out cross validation (LOO-CV; Fig. 1d). After training, the model's external sensitivity and specificity to intent vs. inference was validated on the held-out intent and inference validation sets (see Fig. 1f & Fig. S5).

The model's internal training accuracy is on par with previously published fMRI models of emotion (see Fig. S9)— its average within-subject correlation between predicted and actual ratings across each CV fold is  $r=0.65 \pm 0.34$  (STD; standard deviation) which was significantly greater than zero ( $t(99)=18.91, P<0.001, \text{CI}=[0.58-0.71]$ ; overall Pearson's  $r=0.50, P<0.001$ ; CV mean square error (MSE) = 1.50; Fig. 2a). External validity was verified in the held-out intent validation set by calculating the average of prediction-outcome correlations (average  $r=0.19 \pm 0.002$ ) and then testing if that value was greater than zero in a two-tailed one-sample  $t$ -test ( $t(99)=9.65, P<0.001, \text{CI}=[0.15-0.23]$ , Cohen's  $d=0.23$ ; see Fig. 2b). To test its specificity to intent, we repeated this analysis in the held-out inference validation set. The difference was not significant (average  $r=0.18 \pm 0.002; t(99)=0.48, P=0.629, \text{CI of mean difference}=[-0.06-0.04]$ ) in a two-tailed paired  $t$ -test. Both the training and validation sets are derived from the same observers, but stimuli in the held-out validation sets are unimodal (audio-only and visual-only) while the stimuli in the training sets are multimodal (audiovisual). Full double dissociation was established internally (within the modality of the training data) by performing this validation procedure on fMRI quintile maps derived within each audiovisual stimulus (two-tailed paired  $t(22)=4.63, \text{STD}=0.27, \text{Cohen's } d=1.13, P<0.001$ ; Fig. S10A). Further testing of this model on audiovisual and other socioemotional stimuli is necessary to determine its out-of-sample sensitivity and specificity.

Our model of the target's self-reported emotional intensity is comprised of a distributed pattern of brain activity that relies most significantly on activity in the right visual and anterior insular cortices as well as the right angular gyrus, left posterior cingulate (PCC), bilateral precuneus, and bilateral superior and inferior frontal gyri. These features of importance were determined by conducting a bootstrap hypothesis test (5000 bootstrap samples) over the model weights and applying a significance threshold (FDR  $q<0.05$ ; see Table S2 for a complete list). A more liberal threshold is applied to the data in Fig. 2c for visualization.

Next, we compared the unthresholded brain pattern to other brain maps in the NeuroSynth database<sup>28</sup> to characterize functions associated with this network. The intent pattern is largely unique from previously published activation maps but is most similar to those characterizing brain activity related to resting state, theory of mind, person, social, autobiographical, beliefs, spatial, [scene] construction, speech, and self-referential processing, respectively (Fig. 2d). The unthresholded brain pattern was also compared to well-validated brain-based models of emotion, empathy, and interoception to further gauge its sensitivity and specificity to the intended intensity of a target's social signal (see Methods and Fig. S11). The intent model was dissimilar from emotion models that predict how an observer feels in response to an image or event (PINES<sup>29</sup> cosine similarity = -0.03, maximum cosine similarity = 1; NPS<sup>30</sup> cosine similarity = -0.03; social

rejection<sup>31</sup> cosine similarity = -0.04). However, it was weakly similar to naturalistic models of empathic care (cosine similarity = 0.06) and distress<sup>32</sup> (cosine similarity = 0.10). Overall, this suggests that this brain pattern is not reducible to self-reported feelings of emotional intensity or to empathic processes in general.

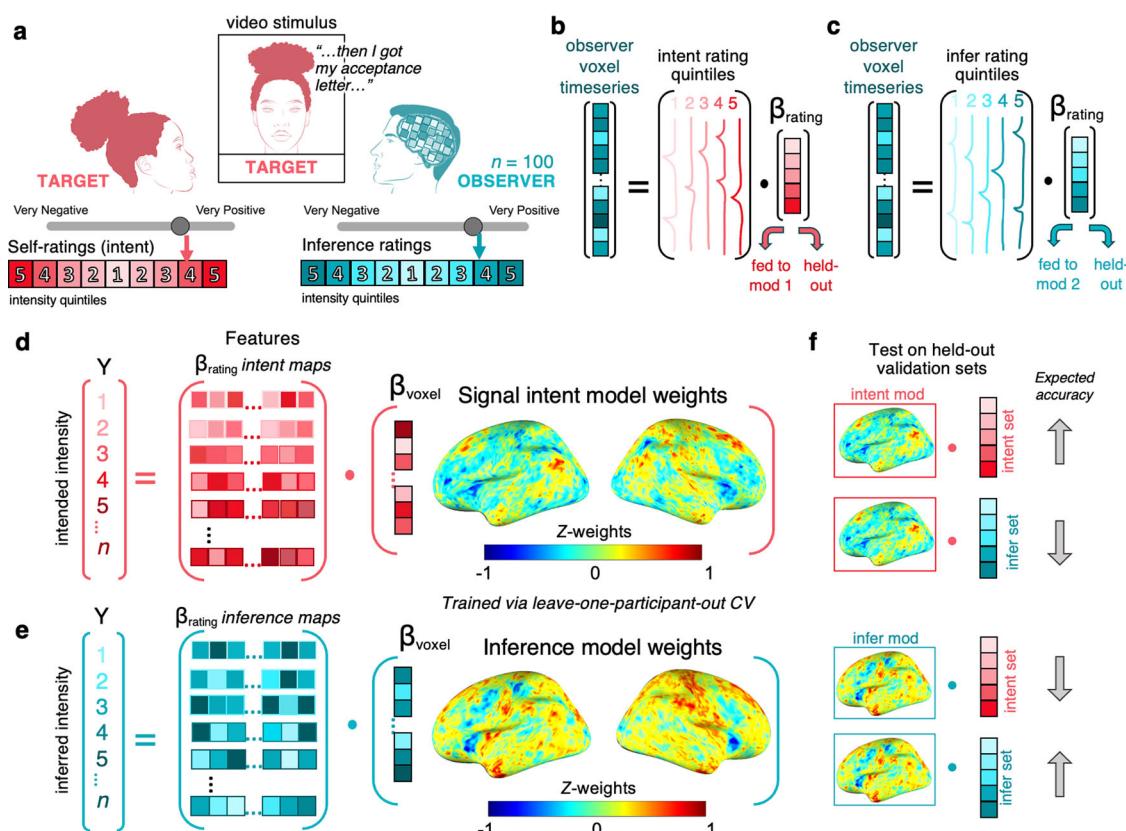
### Socioemotional inference can be predicted from observer brain activity

Using the same procedure described above, we developed a model of brain activity that can predict the level of intensity observers ( $N=100$ ) ascribe to targets on a moment-to-moment basis (Fig. 1e). The model's average within-subject correlation between predicted and actual ratings was  $r=0.68 \pm 0.30$  STD ( $t(99)=22.72, P<0.001$ , CI = [0.62 - 0.74]; overall Pearson's  $r=0.53, P<0.001$ ; CV MSE = 1.52; Fig. 3a). The inference model was verified in the held-out inference validation set (average  $r=0.32 \pm 0.002$ ) and was both sensitive (two-tailed one-sample  $t(99)=12.48, P<0.001$ , CI = [0.27 - 0.37], Cohen's  $d=1.24$ ) and specific to observer inference ratings. That is, its performance on its own validation set was significantly higher than its performance on the

intent validation set ( $r=0.23 \pm 0.002$ ) in a two-tailed paired  $t$ -test ( $t(99)=2.77, P=0.007$ , CI of mean difference = [0.03 - 0.16], Cohen's  $d=0.33$ ; Fig. 3b). To further validate this double dissociation, the model was tested on an additional internal validation set comprised of fMRI quintile maps derived within each audiovisual stimulus (Fig. S10B). Again, the inference model had significantly better performance on the inference test set compared to the intent test set (two-tailed paired  $t(22)=2.08$ , STD = 0.22, Cohen's  $d=0.45, P=0.049$ ), confirming the double dissociation between the two models.

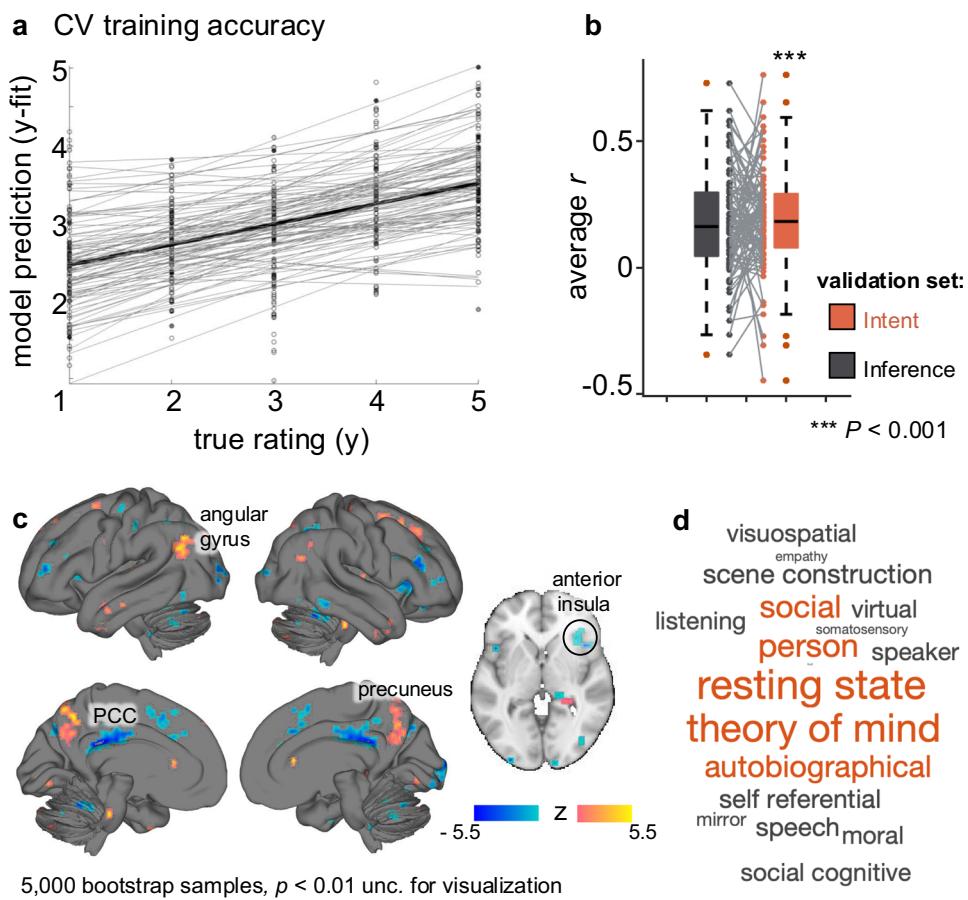
The brain regions that most significantly contributed to the prediction of observers' inferences include the cerebellum (bilateral crus), left precuneus, right primary somatosensory cortex (SI), right inferior frontal gyrus (IFG), bilateral superior medial frontal gyrus, bilateral lingual gyrus, bilateral temporal pole, and bilateral anterior insular cortex (5000 bootstrap samples; FDR corrected  $q<0.05$ ; see Table S3 for complete list; see Fig. 3c thresholded at  $P<0.01$  uncorrected (unc.)).

Again, we compared the unthresholded brain pattern to other published brain maps to characterize functions associated with this



**Fig. 1 | Training schematic for fMRI models of intent and inference.** **a** Targets recorded themselves telling emotionally significant personal stories, then rated their videos on a continuous bivalent scale. Target self-ratings (red) served as the intent of the socioemotional signals conveyed in the videos. Observers viewed these 24 videos in the MRI and rated what they thought the target felt, moment-by-moment, on the same scale. These comprised the inference ratings (blue). There were three types of trials: audiovisual, audio-only, and visual-only; however, the 8 audiovisual trials were the trials of interest (see Fig. S3). Both the intent and inference ratings were transformed into valence-independent intensity quintiles to model the fMRI data. **b, c** Participant-level GLM predicting intent and inference quintiles. A predictor was constructed for each rating quintile ( $q=5$ ) for each stimulus ( $s=24$ ) for each participant ( $N=100$ ) and applied to each participants' voxel timeseries, yielding a set of five whole-brain coefficient maps for each participant, for each model type (intent and inference). Maps from audiovisual trials were used for model training. Maps from unimodal trials were held-out and

used for external validation (see Figs. S4, 5 for details). **d, e** Multivariate model training. Two models were trained from the same audiovisual stimuli: One aimed to characterize signal intent and the other aimed to characterize the observer's inferences. First, brain activity for each intensity quintile, within each participant, was averaged into a single beta map (these voxels comprised the model's features). Next, two multivariate LASSO-PCR models were trained to predict intent and inference intensity quintiles ( $Y=1$  to 5, 5 being the highest intensity) from their corresponding coefficient maps (features) across all participants using leave-one-participant-out cross-validation (LOO-CV). Plotted on the surface maps are the unthresholded normalized predictive Z-weights for each model. **f** External validation. Both models were applied to held-out auditory-only and visual-only trials (see Method detail). The intent model was expected to accurately predict intent ratings, but not inference ratings, and vice versa.



**Fig. 2 | Intent model.** **a** Training. The model was trained to predict the target's self-reported emotional intensity from the observer's brain. A prediction-outcome correlation was calculated across each cross-validated (CV) training fold (average Pearson's  $r = 0.65 \pm 0.34$  STD across folds is significant in a two-tailed one-sample  $t$ -test:  $t(99) = 18.91$ ,  $p < 0.001$ , CI = [0.58 – 0.71]). The target's true self-reported intensity level ( $y$ ) is plotted on the x-axis. The model's prediction of the intensity level ( $y$ -fit) is plotted on the y-axis. **b** Validation. The model was validated by calculating the prediction-outcome correlation across five levels of self-reported intensity within participants in held-out validation trials (see Fig. S5). Average  $r$  values for each participant ( $N = 100$ ) are plotted on the y-axis (orange bar). The model is sensitive to its validation set ( $t(99) = 9.65$ ,  $P < 0.001$ ) in a two-tailed one-sample  $t$ -test. To test its specificity, we repeated this analysis in the inference validation set (gray bar). The intent model had a higher fit on the intent validation

set than the inference set, but the difference was not significant (see Fig. S10A for internal validations). Data are represented in box plots where the median is a black line and the upper and lower whiskers represent the bounds of the quartiles.

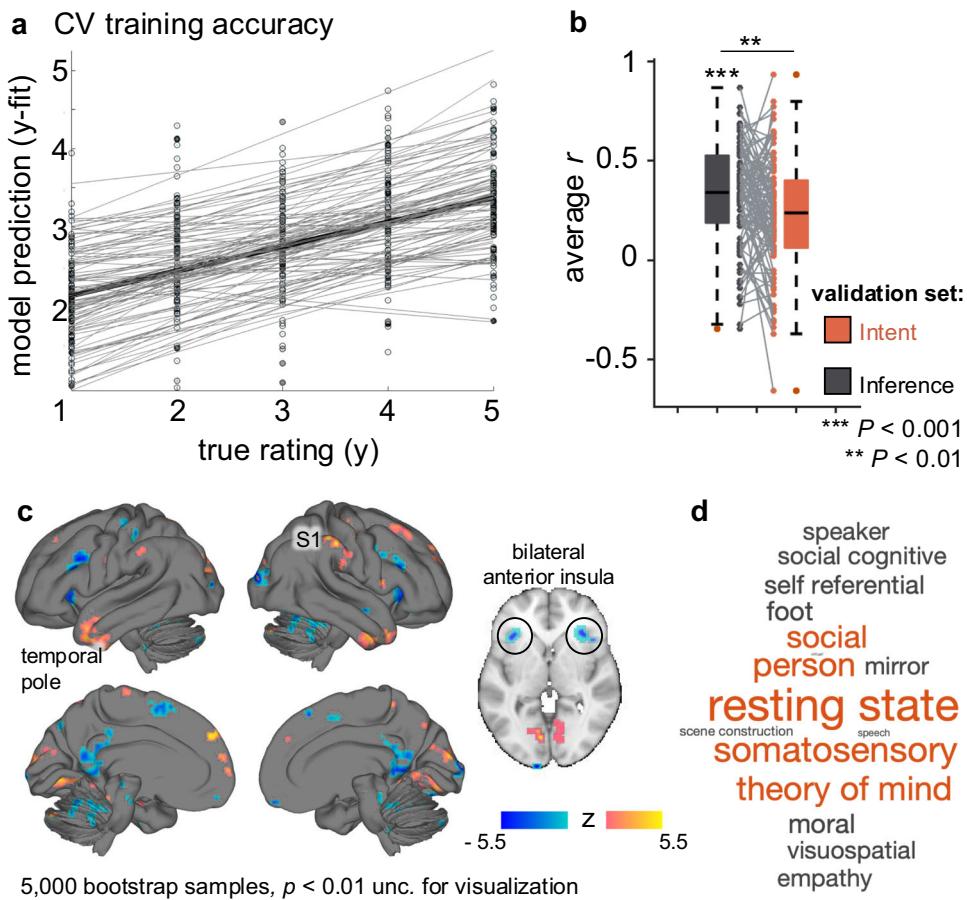
**c** Feature importance. The model's voxel-weight map (Fig. 1b) is loosely thresholded ( $P < 0.01$  uncorrected (unc.)) based on a bootstrap (5000 samples) hypothesis test and plotted for visualization (FDR thresholded regions in Table S1). Activity in right visual and right anterior insular cortices, as well as right angular gyrus, left posterior cingulate cortex (PCC), bilateral precuneus, and bilateral superior frontal gyrus were most important for this prediction. **d** Associated NeuroSynth terms. The unthresholded predictive weight map was fed to the NeuroSynth topic map decoder. A word cloud was constructed of the top 20 terms (excluding singular brain regions). Word size is scaled by strength of similarity. The top five terms are colored orange.

network. The networks and functions in the NeuroSynth database most similar to the inference model were: resting state, somatosensory, person, theory of mind, social, spatial, foot, moral, self-referential, and beliefs, respectively (Fig. 3d). The inference model was unique from the intent model in that it was associated with maps related to somatosensory simulation and bodily action. When we compared the inference model to the brain-based models of emotion, empathy, and interoception in Fig. S11, we found that, like the intent model, the inference model is unique from existing signatures but has a small positive cosine similarity with models of dynamic human-to-human empathetic processing.

**fMRI patterns underlying intent and inference are dissociable**  
Both the intent and inference models were developed on observer brain activity; however, they reveal dissociable components of socio-emotional processing (cosine similarity between intent and inference pattern weights = 0.29). To further investigate their dissociability, we assessed general overlap of voxels that significantly contribute to each model's performance. After thresholding, the only overlap was in the

right anterior insula (Fig. S12), however, these voxel-wise patterns in the bilateral insular cortex (mask from Harvard-Oxford Atlas) were unique (cosine similarity = 0.22). Together, these metrics indicate that the two models capture unique components of socioemotional brain activity.

Next, we tested if subject-level fMRI patterns related to intent and inference are separable at each level of stimulus intensity. To do this, we trained binary linear support vector machines (SVMs) via LOO-CV (see Methods) to separate subject-level intent and inference maps (see Fig. S5) at each level of rating intensity. These are the same maps used as training data for the intent and inference models in Figs. 2 and 3. Each SVMs classification accuracy (acc.) was significantly greater than chance (level 1: acc. =  $70.00\% \pm 3.3\%$ ,  $P < 0.001$ , Area Under the Curve (AUC) = 0.69; level 2: acc. =  $69.00\% \pm 3.3\%$ ,  $P < 0.001$ , AUC = 0.75; level 3: acc. =  $70.00\% \pm 3.3\%$ ,  $P < 0.001$ , AUC = 0.73; level 4: acc. =  $69.00\% \pm 3.3\%$ ,  $P < 0.001$ , AUC = 0.68; level 5: acc. =  $70.00\% \pm 3.3\%$ ,  $P < 0.001$ , AUC = 0.75) indicating that the intent and inference fMRI patterns are linearly separable, and therefore, represent unique processes (Fig. 4a). To better understand the brain regions that distinguish between intent



**Fig. 3 | Inference Model.** **a** Training. The model was trained to predict an observer's inference from their brain activity. A prediction-outcome correlation was calculated across each training fold (average Pearson's  $r = 0.68 \pm 0.30$  STD across folds is significantly greater than zero in a two-tailed one-sample  $t(99) = 22.72, p < 0.001, CI = [0.62 - 0.74]$ ). The observer's true inferred intensity level ( $y$ ) is plotted on the x-axis. The model's prediction of the intensity level ( $y$ -fit) is plotted on the y-axis. **b** Validation. The inference model was verified in its held-out validation trials (average  $r = 0.32 \pm 0.002$  is significantly greater than zero in a two-tailed one-sample  $t(99) = 12.48, P < 0.001, CI = [0.27 - 0.37]$ ). Average  $r$  values for each participant ( $N = 100$ ) in the inference validation set are plotted on the y-axis (gray bar). To test its specificity, we repeated this test in the intent validation set (orange bar). The inference model fits its own validation set better than the intent validation set

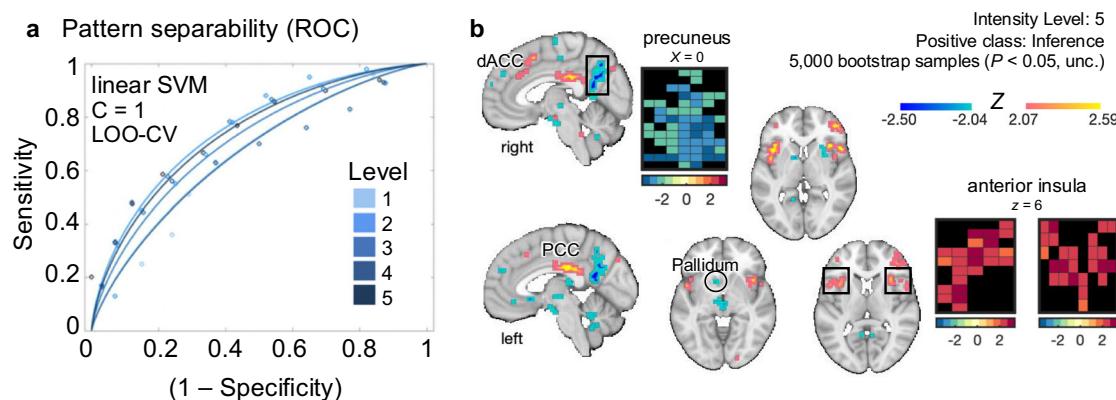
(two-tailed paired  $t$ -test ( $t(99) = 2.77, P = 0.007, CI = [0.03 - 0.16]$ , Cohen's  $d = 0.33$ ). Data are represented in box plots where the median is a black line and the upper and lower whiskers represent the bounds of the quartiles. **c** Feature importance. The voxel-weight map is liberally thresholded ( $P < 0.01$  unc.; see Fig. 1c) based on a bootstrap hypothesis test (5000 samples) and plotted for visualization (see Table S2 for FDR thresholded regions). Activity in the cerebellum (bilateral crus), precuneus, primary somatosensory cortex (S1), inferior frontal gyrus, bilateral superior medial frontal gyrus, lingual gyrus, temporal pole, and anterior insular cortex were most important for this prediction. **d** Associated NeuroSynth terms. The terms in the Neurosynth database most similar to the inference model are plotted as a word cloud and scaled by strength of similarity. The top five terms are colored orange.

and inference, we performed a bootstrap (5000 samples) hypothesis test over the predictive voxel weights for the classifier trained at the highest level of intensity (level 5) and thresholded the weights ( $P < 0.05$ , unc.) so that the brain regions where intent and inference maximally diverge could be compared (Fig. 4b). Patterns of activity in the dorsal anterior cingulate (dACC), PCC, anterior insula, pallidum, and precuneus maximally separate intent and inference at the highest level of socioemotional intensity. These regions correspond with those that diverge at the lowest level of intensity (see Fig. S13). Together, these analyses indicate that the multivariate patterns which comprise the intent and inference models are dissociable from one another and therefore reflect unique, but simultaneous, components of socioemotional processing.

#### Alignment of intent and inference is related to empathic accuracy

After we identified these two components of socioemotional processing, we sought to test how they interact in relation to an individual person's empathic accuracy. To do this, we applied the intent and inference models to participant-level fMRI data when participants

made (a) inaccurate inferences (low empathic accuracy) and (b) accurate inferences (high empathic accuracy; see Fig. 5a, Fig. S5 and S15–16, and Methods for details). Then we correlated the predictions of the two models across all participants. When participants are inaccurate, there is more variance across the predictions of the intent and inference models, and, therefore, they are weakly correlated ( $r = 0.28, P = 0.004$ ; Fig. 5a). However, when participants are highly accurate, there is less variance between the model predictions, and they are better correlated ( $r = 0.64, P < 0.001$ ). The alignment between the two models was significantly greater during high accuracy performance than low accuracy performance (two-tailed  $z$ -test of the correlation difference  $z = 3.26, P = 0.001, CI_{r1} = [0.51 - 0.74], CI_{r2} = [0.09 - 0.45]$ , Cohen's  $q = 0.47$ ; Fig. 5b). We verified this effect in the validation trials (low empathic accuracy alignment  $r = 0.58, P < 0.001$ ; high empathic accuracy alignment  $r = 0.79, P < 0.001$ ; two-tailed  $z = 2.83, P = 0.005, CI_{r1} = [0.70 - 0.85], CI_{r2} = [0.43 - 0.70]$ , Cohen's  $q = 0.41$ ; Fig. 5c). This analysis indicates that when an observer is making accurate inferences, both neural patterns predictive of signal intent and neural patterns predictive of observer inference are highly similar to the observer's brain activity at that moment. That is, greater



**Fig. 4 | Separability of intent and inference fMRI patterns.** **a** We tested if brain activity patterns ( $N=100$ ) related to the target's self-report (signal intent) and the observer's inference are separable at each level of stimulus intensity. Binary SVMs revealed pattern separability at each level (level 1: accuracy (acc.) =  $70.00\% \pm 3.3\%$ ,  $P < 0.001$ , AUC = 0.69; level 2: acc. =  $69.00\% \pm 3.3\%$ ,  $P < 0.001$ , AUC = 0.75; level 3: acc. =  $70.00\% \pm 3.3\%$ ,  $P < 0.001$ , AUC = 0.73; level 4: acc. =  $69.00\% \pm 3.3\%$ ,  $P < 0.001$ , AUC = 0.68; level 5: acc. =  $70.00\% \pm 3.3\%$ ,  $P < 0.001$ , AUC = 0.75). Receiver operating curves (ROC) visualize the sensitivity and specificity of the SVMs which separated the patterns. Curves are colored shades of blue, where lighter shades are lower

intensity levels and darker shades are higher levels. **b** To better understand the brain regions that distinguish between signal intent and inference, the predictive voxel weights for the SVM classification at level five (high intensity) a bootstrap hypothesis test (5000 samples) was conducted over the weights and a weak significance threshold ( $P < 0.05$ , unc.) was applied. Here, inference maps are the positive class. Patterns of activity in the dorsal anterior cingulate (dACC), posterior cingulate (PCC), pallidum, insula, and precuneus maximally separate intent and inference at the highest level of socioemotional intensity (see Fig. S12 for maps of the other levels).

concordance between the neural pattern predictive of target's intended signal intensity and the neural pattern predictive of observer inferences indicates greater empathic accuracy. Quantitative descriptions of observer accuracy across stimuli in this study are included in Figs. S7 & S8.

**Exploratory: brain regions that transform intent into inference**  
 In a post-hoc exploratory univariate analysis, brain activity during low accuracy trials (see schematic in Fig. 5a) was used to predict (X) the pattern expression of the inference model (Y) while controlling for the pattern expression of the intent model. Pattern expression is the dot-product between two multivariate patterns (see Methods for details). Low accuracy trials were selected so regions where inference-related activity maximally diverged from activity related to intent recognition could be isolated. The resulting univariate model weights were corrected for multiple comparisons (FDR  $q < 0.05$ ,  $k = 25$ ). Positive clusters were revealed in the right primary somatosensory cortex (S1) and right parahippocampal gyrus (PHG; Fig. 6). Negative clusters were revealed in the left insula and left primary motor cortex (M1). Observers tap their fingers when they make a rating, therefore, we tested if this motor activity was similar to activity related to finger tapping in the NeuroSynth Database. It was dissociable (*cosine similarity* = -0.07) and therefore, unlikely to be related to the execution of button pressing. These exploratory results may suggest that S1, M1, and PHG are associated with the formation of socioemotional inferences above and beyond latent recognition of a signal's intent. Albeit, in this analysis, activity in these regions is implicated in the formation of an incorrect inference. Note, though individual observer's accuracy in this task varies (see Figs. S7 and S8A), observers' own emotional intensity in response to the target (i.e., how the observer feels) was consistently correlated with what they inferred the target was feeling (average within-person  $r = 0.89$  (0.19 STD); see Fig. S8B). This suggests that when an observer makes an incorrect inference about a target, they may be reporting what they feel themselves and may use autobiographical memory to do so.

#### Exploratory: functional connectivity predictive of inference accuracy

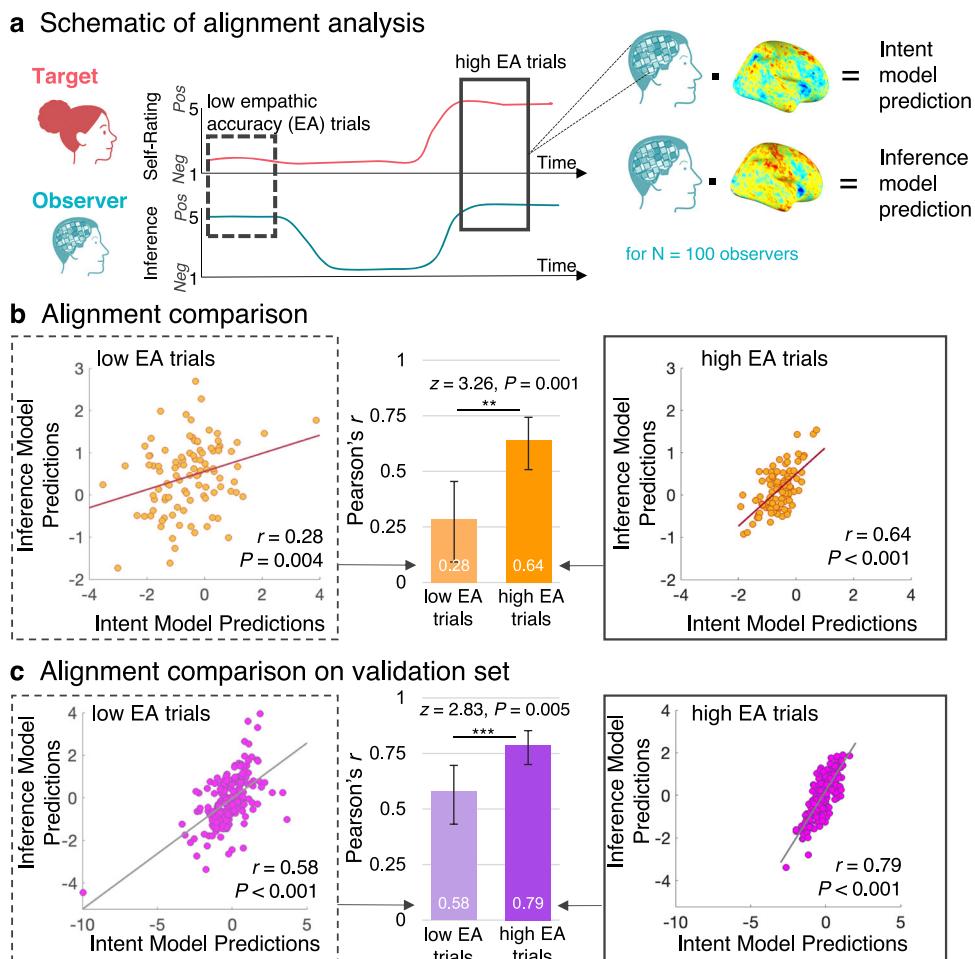
Pairwise distance correlation matrices were calculated across pre-processed timeseries data from 272 brain regions from the

Brainnetome Atlas<sup>33</sup> for each audiovisual stimulus (Fig. S17A, B). Degree centrality was calculated for each node, on average, for each participant and then correlated with participant's overall empathic accuracy performance (Pearson's  $r$ ) on the audiovisual portion of the task (Fig. S17C–E). Degree centrality is equal to the number of ties a node has. High degree centrality indicates high importance of a node in a connectivity network. We found three significant correlations ( $P < 0.05$ , unc.) in this exploratory post hoc analysis: right PHG ( $r = 0.22$ ,  $P = 0.032$ ), right cingulate gyrus (rostral ventral area 24;  $r = 0.24$ ,  $P = 0.019$ ), and (e) right inferior temporal gyrus (ITG;  $r = 0.22$ ,  $P = 0.032$ ). This suggests that activity in these brain regions is related to empathic accuracy.

## Discussion

The effectiveness of social signaling depends upon both the intended meaning of the signal and the interpretation of the signal by an observer (i.e., socioemotional inference)<sup>6</sup>. To better understand how social signaling is processed in the brain, we sought to dissociate the neural patterns underlying signal intent and inference in a dynamic naturalistic storytelling paradigm and test how these patterns relate to empathic accuracy. We found that both intent (the target's self-reported emotional intensity) and observers' inferences about the target's emotional intensity could be predicted from observers' brain activity. The multivariate brain patterns derived from these predictions are dissociable; however, when the models' predictions align, observers make more accurate inferences. We interpret these findings to suggest, in part, that there is some latent recognition of a social signal's intended intensity that observers can draw upon when forming a conscious inference. Moreover, these findings suggest that observers can have an accurate latent representation of another person's intent even when making an incorrect inference.

How can it be possible that the intended intensity of another person's social signals can be decoded from observers' brain activity, and that this neural pattern is dissociable from the pattern predictive of observers' conscious inferences? Human adults have highly adept schemas for social information that are activated when they perceive prototypical socioemotional expressions<sup>34</sup>. Prior work has found that rich, category-specific visual features can be readily mapped to distinct emotions that are coded across many brain regions, including primary visual cortex<sup>21</sup>. Unless a target is intentionally trying to deceive an



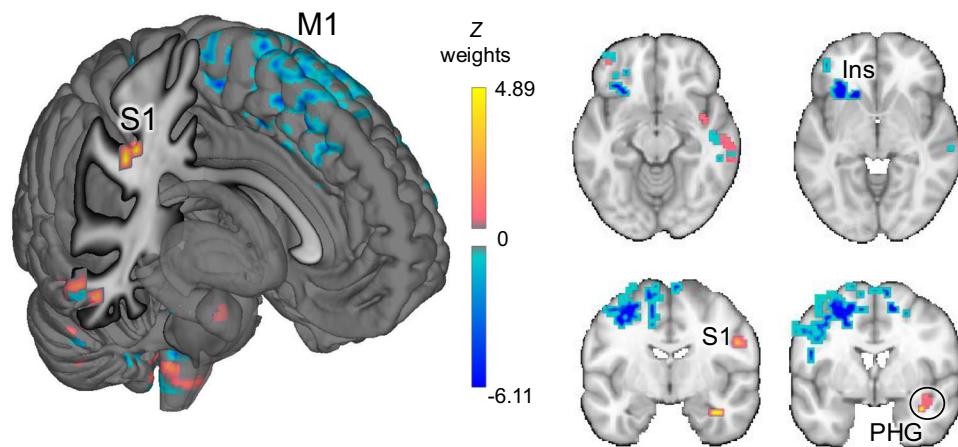
**Fig. 5 | Alignment between the intent and inference models is related to empathic accuracy.** **a** Schematic of alignment analysis. To test how each model's predictions relate to an observer's empathic accuracy, we took the dot-product between each model and each individual subject's ( $N=100$ ) brain activity during periods of time when observer empathic accuracy performance was at its highest and lowest (see Fig. S5), and then added in the models' intercepts. **b** Alignment comparison. To test how the alignment of the intent and inference patterns are related to empathic accuracy, we correlated each model's predictions across all participants. When participants are inaccurate, there is more variance across the predictions of the intent and inference models, and therefore they are weakly correlated ( $r=0.28, P=0.004$ ). However, when participants are highly accurate, there is lower variance between the models' predictions, and they are better correlated ( $r=0.64, P<0.001$ ). The alignment between the two models was

significantly greater during high accuracy performance than low accuracy performance (two-tailed Fisher's  $z$ -test of the correlation difference  $z=3.26, P=0.001$ ,  $CI_{r1}=[0.51-0.74]$ ,  $CI_{r2}=[0.09-0.45]$ , Cohen's  $q=0.47$ ). Both  $r$  values from the scatter plots are also plotted as orange bars with confidence intervals derived from Fisher's  $z$ -tests. **c** Alignment comparison on validation set. This result—that alignment between the two models was significantly greater during high accuracy performance ( $r=0.79, P<0.001$ ) than low accuracy performance ( $r=0.79, P<0.001$ ) was replicated in held-out validation trials (two-tailed Fisher's  $z=2.83, P=0.005$ ,  $CI_{r1}=[0.70-0.85]$ ,  $CI_{r2}=[0.43-0.70]$ , Cohen's  $q=0.41$ ). The  $r$  values from the validations are plotted as purple bars with confidence intervals derived from the Fisher's  $z$ -tests. This analysis indicates that the models developed in Figs. 2 & 3 can be combined to predict an observer's empathic accuracy.

observer, or has a disorder that impacts socioemotional communication, a target will convey information in a manner that will activate the correct schema in an observer<sup>14</sup>. That is the function of social signaling—to convey information in a manner that will be quickly and accurately understood by the individual an animal is conveying it to<sup>6</sup>. We suspect that the intent pattern revealed in this investigation is capturing this process of schema-activation in the observers. Schema activation in this investigation is specific to the intensity of a signal.

Subcortical brain regions commonly implicated in emotion perception (i.e., the amygdala, striatum, and periaqueductal gray) were not identified among the features of highest importance in the intent or inference models. Instead, the distributed patterns of important features primarily included cortical regions that perform complex multisensory integration (i.e., angular gyrus, frontal gyrus, PCC, temporal pole, and the precuneus)<sup>35</sup>. From a constructionist perspective, which theorizes that emotions are flexible processes inseparable from

the context in which they emerge<sup>36,37</sup>—the intent pattern may reflect neural processing that abstracts the perceptual input (i.e., facial expressions, body movements, speech, and vocal intonations) into socioemotional schemas related to intensity. Similarly, the inference model may represent a subsequent stage of processing, where the information the target signaled is related to the observers' autobiographical past, current experiences, and their expectations for the future. Indeed, observers' overall inferences were highly correlated with their own emotional responses to the stories (Fig. S8C), suggesting that autobiographical recall and mood induction are concurrent, if not interwoven, processes contributing to inference formation. The voxel pattern of the inference model, which relies primarily on mentalizing networks and brain areas implicated in social abstraction and somatosensory processing (the temporal pole and S1, respectively)<sup>38</sup>, provides further support for this hypothesis. Being that the intent and inference models were (1) dissociable, (2) verifiable in their own held-out validation sets, and (3) unique from other published



**Fig. 6 | Exploratory analysis of brain regions uniquely related to the inference model's pattern expression during low accuracy trials.** In this exploratory univariate analysis, brain activity during low accuracy trials (see schematic in Fig. 5 and Fig. S4) was used to predict the pattern expression of the inference model while controlling for the pattern expression of the intent model across participants ( $N = 100$ ). Plotted are the standardized weights corrected voxel-wise ( $q < 0.05$  FDR) for multiple comparisons. Clusters of 25 contiguous voxels were retained for display and interpretation purposes. Warm colors indicate positive weights and cool

colors indicate negative. When controlling for expression of the intent pattern, activity in the right primary somatosensory cortex (S1) and parahippocampal gyrus (PHG) increases with the inference pattern expression, while activity in the left insula (Ins) and primary motor cortex (M1) decreases with the inference pattern expression. This analysis captures brain regions that may influence the inference pattern (brain activity predictive of an observer's inference of socioemotional intensity), independent of the intent pattern.

models of emotion induction, like PINES<sup>29</sup>, we suspect that they capture two unique components of social signal processing—schema activation and deliberate inference formation.

When the intent and inference models' predictions aligned in the brains of the observers, observers made more accurate inferences. Furthermore, the two models could be combined to predict the empathetic accuracy in the held-out validation sets. Our follow-up exploratory analysis indicated that when individuals made inaccurate inferences, activity in right S1 and PHG increased with the intensity of their inferences, above and beyond the expression of the intent pattern. The centrality of the PHG in a network of functional connectivity during the inference task was also correlated with individual differences in overall empathetic accuracy performance (Fig. S17). We speculate that somatosensory simulations, via S1, and autobiographical recall, via PHG<sup>39</sup>, support the transformation of a socioemotional signal into a conscious, reportable inference. However, interpretations of this activity should be approached with caution due to the exploratory nature of the analysis. Part of the transformation from intent to inference in this paradigm requires a motor response: Observer's must update their ratings with a button press. Therefore, it is difficult to disentangle this brain pattern from button pressing entirely; however, this pattern of brain activity was dissimilar from patterns associated with finger tapping in the NeuroSynth database. Furthermore, though the left lateralization of activation in M1 was consistent with the (right) hand making the ratings, the positive clusters in S1 and the PHG were ipsilateral to the ratings-hand. Prior research that shows S1 activates during motor imagery and empathy<sup>40–43</sup>. Furthermore, a growing thrust of empathy research purports a role for somatosensory simulations in the understanding of social interactions<sup>44–46</sup>. Together, these results suggest that inference involves an internalizing of the events described by the target, and that people simulate the actions described and relate them to their own prior experiences and expectations. This summary dovetails with our behavioral finding that observers consistently report their own feelings to be concordant with what they think the target feels (Fig. S8). That is, participants draw upon their own experiences and feelings when trying to make sense of the feelings of others<sup>47</sup>.

This study has several limitations. First, intent and inference ratings occur spontaneously and simultaneously in this paradigm which

makes it difficult to ensure complete separation of these constructs in the observer's brain. We did, however, perform independent validations which supported the dissociability of the intent and inference models across stimuli (see Fig. S10). Second, target self-reports may not perfectly match their actual feeling state at the time of story; it could be influenced by self-reflection, mood at the time of report, or by social desirability<sup>48</sup>. Third, the generalizability of these findings to different cultures is unknown and, due to cultural variation in verbal and non-verbal communication, it is possible that our sample, which is largely comprised of Stanford University undergraduates, cannot be reproduced in other age, socioeconomic, or cultural groups (see Fig. S1). Fourth, though the stimuli themselves are highly dynamic and complex, the models are trained to predict only a single dimension of socioemotional information: intensity. This was done because naturalistic stories often signal positive and negative information at a faster rate than we can sample fMRI data<sup>49</sup>. For example, a participant may be describing both the sadness and the love they felt after the death of a family member. These are intense, complex emotions, therefore, removing valence from individual ratings allowed us to better model dynamic shifts in emotion signaling and to isolate signatures of signal intent from observer inference. Indeed, including valence information did not improve the accuracy of our models (see Fig. S13). Further validation of our models on other naturalistic and social audiovisual data is necessary to determine their sensitivity and specificity to both a target's self-reported internal emotional state and an observer's conscious inference of that state. Additionally, we did not analyze story content or language in this study because video randomization reduced our power to isolate content-specific activity (not every participant observed the same audiovisual story). Future work will seek to train finer grained models of story context over time to decode higher-dimensional socioemotional content from observer brain activity.

Various neuroimaging studies have attempted to predict aspects of socioemotional processing from human brain activity (see Fig. S9 for a summary). Here, we situate socioemotional processing within the ethological framework of social signaling and inference. This framework is important—Sharing information is essential to the well-being of individuals and their communities because individuals must interact with each other to achieve personal needs that cannot be achieved

alone. Effective signaling can engender social bonds, mutual aid, and collaboration<sup>50</sup>. Ineffective signaling, however, can be costly. A missed alarm call can result in death, while misunderstanding social signals can result in ostracization or rejection. By providing insight into how the human brain interprets social signals, the current research may provide inspiration for the development of interventions aimed at reducing loneliness and social isolation.

## Methods

### Experimental model and subject details

One hundred (59 Women, 37 Men, 4 No Response, average age = 25.23 STD = 9.96) adult healthy members of the Stanford University community participated in this study. Gender was determined via participant self-report. Participants were asked to indicate which gender they best identified with (male or female) and were given the option to not identify. Forty-two participants identified as White, 32 as Asian American, 14 as Hispanic/Latinx, 7 as Black, and 5 did not report their race (see Supplementary Fig. 1a). Participants were asked to provide their gender in binary terms (male or female). Participants were asked to complete the MacArthur Scale of Subjective Social Status<sup>51</sup>. Average perceived socioeconomic status (SES) on the 10-point ladder scale was 6.71 (STD = 1.57). SES was skewed towards the upper and upper-middle classes (Mode = 7; Median = 7). Eighty-nine participants reported having at least some college education (this includes those currently enrolled as Stanford undergraduates). Eighty-three reported that their parents had at least some college education. Participants were all right-handed.

Participants were recruited through internal and surrounding communities near Stanford University in Stanford, CA. All participants gave informed consent and were compensated monetarily via Amazon gift cards for their participation. Though we aimed to have a representative sample in both the target sample and observer sample, neither sex nor gender was a part of the study design. Effects of gender were not expected or intended to be studied. A post-hoc analysis concluded there was no effect of gender on inference accuracy (Table S5). This study was approved by the Stanford University Institutional Review Board.

### Experimental procedures

**Eligibility criteria.** Participants were required to be between the ages of 18 to 65. Participants who had contraindications for the MRI environment were excluded from this study. Sample size was determined via previously published studies of empathic accuracy and narrative storytelling<sup>23</sup>. No statistical method was used to predetermine sample size.

**Stimuli.** We selected a subset of 24 videos (19 unique targets or storytellers) from a curated high-quality video dataset of 193 video clips of 49 volunteers describing emotional life events known as the Stanford Emotional Narratives Dataset (SENDv1)<sup>22</sup>. The targets used in this selection were balanced for gender (11 Women and 8 Men) and racial representation (9 White, 5 Black, 4 Asian, and 1 Latinx/Hispanic; see Fig. S1A). Many of the targets were current students. Self-ratings of bi-valent emotional intensity were obtained from targets on a moment-by-moment bipolar scale (very negative to very positive). Targets recorded their videos and then immediately watched them to provide the self-ratings. Twelve of the 24 videos featured negative life events, while the other 12 featured positive life events (see Fig. S2 for a visual summary of story content). Videos were cut down in length so that they ranged from 1 - 3 minutes. Five targets contributed two separate videos (one positive, one negative) to this dataset, however, trials with duplicate targets were not used in the main analysis (see Experimental Session for more information). Stimuli were displayed using Psychtoolbox v3 via MATLAB R2017B.

**Randomization.** This is a within-subjects design, where stimulus presentation was pseudorandomized in three a priori orders. Each subject was randomly assigned to an order when they signed up for the experiment. Video presentation and intertrial interval (ITI) lengths were shuffled and fixed to an order number. The Investigators were not blinded to allocation during experiments and outcome assessment.

**Practice session.** All participants completed a “practice session” before the experiment, where they learned how to make their ratings on the slider provided. Participants viewed 10-s clips of stimuli from the SENDv1 dataset that are unique from those used in the study.

**Experimental session.** All 100 participants were shown the narrative story stimuli while inside the fMRI environment. While viewing, participants rated the emotional-narrative stimuli moment-by-moment on the rating scale by way of button presses that moved a slider on screen. Participants pressed the buttons to move it right or left with their index and middle finger, respectively, of their right hand. There were three sensory conditions in this experiment: a visual-only condition, where participants could see but not hear the videos (8 trials); an auditory-only condition where they could hear but not see the videos (8 trials); and an audiovisual condition where they could both see and hear the videos (8 trials). There were 24 trials total, divided into two runs of 12 trials each (See Fig. S3A). The audiovisual condition (8 trials) was the basis of the model training data and is the modality of interest in the present study. The other two conditions were combined and used as a held-out validation set. Due to a coding error, only 23 videos were spread across the audiovisual condition during counterbalancing (the missing video was negatively valanced; see Fig. S7 for a depiction of the valence of all the audiovisual stimuli). Five targets contributed two separate videos (one positive, one negative) to this dataset, however, the sensory channel of presentation was not repeated within targets. Videos were separated by an ITI (fixation cross) that ranged from 4 - 12 s.

**Post-experiment session.** Immediately after the fMRI experiment, participants were shown 10 s clips of each video they watched in the scanner, in the same sensory modality (auditory-only, visual-only, and audiovisual) and order that was presented to them in the scanner. First, participants answered True-False questions about the story to assess comprehension and attention. Next, participants were asked if the target was familiar to them. The targets were recruited on Stanford campus; therefore, it was important to confirm that our participants were not friends or acquaintances with the targets. Participants were also asked to rate the overall emotional intensity of the stimulus as well as the overall emotional intensity they felt in response to the stimulus on the same bi-valent scale they used during the experiment. This session was displayed via MATLAB R2017B on an Apple laptop computer in an experiment room in the lab.

### Data collection

**Ratings acquisition.** Participants were asked to “rate how positive or negative you think the target is feeling moment to moment” on a bipolar visual analog scale (VAS) with anchor points very negative to very positive. Participants did not see any number responses, but their ratings were recorded on a scale from 0 to 100. Ratings were collected via a 2 × 4 bimanual button box through the FORP 932 modular response box system. Ratings were collected through Psychtoolbox v3 (<http://psychtoolbox.org/>) in MATLAB R2017B, and were sampled by Psychtoolbox every 0.5 s. One button press moved the slider by 12 pts. The slider could be moved to the right or left via buttons pressed by the index and middle finger on the participant’s right hand. See Supplementary Fig. 4 for an example of raw ratings time courses for one video.

**Neuroimaging acquisition.** The study was conducted at the Stanford University Center for Cognitive and Neurobiological Imaging (CNI) using a 3 T GE Discovery MR750 scanner (running ESE version 23.1 v02: the GE operating system) and a 32-channel Nova Medical head coil. The scan began with a 3-plane localizer or scout scan to find the participant's head. Next, we collected a 3D T1-weighted anatomical scan at 0.9 mm isotropic resolution (Flip Angle = 12; FOV = 23). This scan uses GE's "BRAVO" sequence. This was followed by a higher-order shim and a field map was collected before each functional scan. Functional scans used a gradient echo sequence (Full brain BOLD EPI, TR = 2 s, TE = 25 ms, FOV = 23.2, Flip Angle = 77, 46 slices at a 2.9 mm slice thickness, Acquisition Order = Interleaved, voxel size = 2.9 mm<sup>3</sup>). A 3x in-plane acceleration was used to reduce EPI distortion. Functional image acquisition was divided into two runs. Between runs there was a break of approximately 15–30 s where the experimenter checked the participant to make sure they were comfortable and alert.

### Quantification and statistical analysis

**Outlier removal.** No outlier subjects were found nor removed from this dataset. One participant experienced a technical error where they were shown duplicate trials. Duplicate trials were removed from this participant's data prior to analysis. No other data were excluded from the analyses.

**Preprocessing of ratings.** Intent and inference ratings were down-sampled from (0.5 s) to the rate of the TR (2 s). Next, both sets of ratings were range-normalized within participants (for observers, range normalization was done across all stimuli). Range normalization involved subtracting the minimum of that participant's ratings then dividing by the [max - min] of that person's ratings, from a rating at any given time point. This is also known as feature scaling, and it yields values that are shifted and rescaled so that all values range between 0 and 1. Next, those ratings were adjusted into 5 levels of valence-independent emotional intensity. That is, ratings between 0.41–0.50 and 0.51–0.60 were set to level 1; ratings between 0.31–0.40 and 0.61–0.70 to level 2; ratings between 0.21–0.30 and 0.71–0.80 to level 3; ratings between 0.11–0.20 and 0.81–0.90 to level 4; and ratings between 0–0.10 and 0.91–1.00 to level 5. Level 5 is the highest intensity rating. This was done because naturalistic stories often combine positive and negative emotions and can signal positive and negative information at a faster rate than we can sample the brain data. For example, a participant may be describing both the sadness they felt and the love they felt after the death of a family member. These are both intense, complex emotions, so removing valence from individual ratings allows us to better model dynamic shifts in emotion signaling and to isolate signatures of signal *intent* from an observer's *inference*. This decision was inspired by the Circumplex Model of Emotion<sup>52</sup>. However, for completeness, we include an analysis that attempts to predict five levels of valence (both intent and inference) from observer brain activity (see Fig. S11).

**Derivation of empathic accuracy.** Empathic accuracy ratings were constructed by first subtracting the normalized intent ratings from the normalized observer inference ratings, across the whole time series. These ratings were then converted into five levels where the highest (level 5) reflects times where there was no difference between the observer's ratings and target intent, that is, when observers were most accurate. The lowest (level 1) reflects when there was a maximal difference between the observer's ratings and target intent in either direction; that is, when observers were least accurate. This measure is not valence independent. For additional video-by-video metrics of moment-by-moment empathic accuracy that confirm task validity see Supplementary Table 1.

**fMRI preprocessing.** Imaging data were first converted into BIDS via in-house scripts. All data were then preprocessed using FMRIPREP version 1.4.1-2019<sup>53,54</sup> a Nipype based tool<sup>55,56</sup> (RRID: SCR\_002502) based tool. Each T1w (T1-weighted) volume was corrected for INU (intensity non-uniformity) using N4BiasFieldCorrection v2.1.0<sup>57</sup> and skull-stripped using antsBrainExtraction.sh v2.1.0 (using the OASIS template). Brain surfaces were reconstructed using recon-all from FreeSurfer v6.0.1<sup>58</sup> (RRID: SCR\_001847), and the brain mask estimated was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray matter of Mindboggle<sup>59</sup> (RRID: SCR\_002438). Spatial normalization to the ICBM 152 Nonlinear Asymmetrical template version 2009<sup>60</sup> (RRID: SCR\_008796) was performed through nonlinear registration with the antsRegistration tool of ANTs v2.1.0<sup>61</sup> (RRID: SCR\_004757), using brain-extracted versions of both T1w volume and template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast<sup>62</sup> (FSL v5.0.9, RRID: SCR\_002823).

Functional data was slice time corrected using 3dTshift from AFNI v16.2.0<sup>63</sup> (RRID: SCR\_005927) and motion corrected using mcflirt (FSL v5.0.9)<sup>64</sup>. This was followed by co-registration to the corresponding T1w using boundary-based registration<sup>65</sup> with six degrees of freedom, using bbregister (FreeSurfer v6.0.1). Motion correcting transformations, BOLD-to-T1w transformation and T1w-to-template (MNI) warp were concatenated and applied in a single step using antsApplyTransforms (ANTS v2.1.0) using Lanczos interpolation. Physiological noise regressors were extracted applying CompCor<sup>66</sup>. Principal components were estimated for the two CompCor variants: temporal (tCompCor) and anatomical (aCompCor). A mask to exclude signal with cortical origin was obtained by eroding the brain mask, ensuring it only contained subcortical structures. Six tCompCor components were then calculated, including only the top 5% variable voxels within that subcortical mask. For aCompCor, six components were calculated within the intersection of the subcortical mask and the union of CSF and WM masks calculated in T1w space, after their projection to the native space of each functional run. Framewise displacement<sup>67</sup> was calculated for each functional run using the implementation of Nipype.

**Univariate analysis.** Preprocessed BOLD runs were concatenated for each participant. Next a participant-specific design matrix was created. It included 36 nuisance regressors extracted via FMRIPREP. These nuisance regressors included CSF and white matter regressors and their derivatives as well as 3D motion regressors and their derivatives, and spikes or motion outliers. Next, the regressors of interest were added to the design matrix and three unique "single-trial" models were fit to the observers' brains. The first modeled "signal intent," the next "observer inference," and the last modeled "empathic accuracy."

**Description of training and validation sets.** Regressors were constructed for each rating quintile for each video stimulus (24 trials) and input into the participant's design matrix (see Fig. S3B). 120 is the max number of ratings regressors an individual participant could have. Because participants made individual ratings for each video, the number of regressors for each participant varied. Participants made on average 96.07 (10.97 STD; range = [64 – 115]) ratings each, across the whole study. Ratings regressors were convolved using the canonical HRF function from the SPM12 toolbox (<https://www.fil.ion.ucl.ac.uk/spm/>). The nuisance regressors and intercept were added to each design matrix and then an ordinary least squares (OLS) regression was run for each participant. This entire procedure was repeated three times, for the three unique rating types (target intent self-ratings, observer inference ratings, and empathic accuracy). Variance Inflation Factors (VIF) were calculated for all regressors of interest to

perform data quality control. Event trials with a VIF greater than three standard deviations from an individual's mean were removed as outliers. We tested for intercorrelation across the design matrix and found no issues of multicollinearity within the regressors of interest.

The results of these participant-level single-trial univariate models of the fMRI timeseries data were six sets (five quintiles) of coefficient maps for each observer. The first set contained five levels of "intended" intensity for audiovisual stimuli. This became the training set for LASSO-PCR model 1 (the intent model featured in Fig. 2). The second set contained five levels of inference intensity for audiovisual stimuli. This becomes the training set for LASSO-PCR model 2 (the inference model featured in Fig. 3). The separability of these two sets of participant-level data was later tested via SVM (see Fig. 4). The validation data for models 1 and 2 are comprised of the third and fourth sets of data, respectively, which contain five levels of intended intensity for auditory-only and visual-only stimuli (validation set for model 1) and five levels of inference intensity for auditory-only and visual-only stimuli (validation set for model 2). Finally, the fifth and sixth sets contain five levels of empathic accuracy. The fifth set is for audiovisual stimuli and the sixth set is for auditory-only and visual-only stimuli. These empathic accuracy models are agnostic to level of intensity. The behavior of models 1 and 2 are tested on these unique data in Fig. 5.

### Predictive analyses

**Model training.** Whole brain coefficient maps derived from the single-trial models were averaged within each intensity quintile so that each participant had five coefficient maps, one for each level of intensity. We implemented a LASSO-PCR<sup>27</sup> model (error type = Mean Squared Error (MSE) where the number of principal components (PCs) was automatically determined using the rank of the scaled training data matrix (rank function in MATLAB). Voxel weight maps were generated by projecting a linear combination of all PCs' weights back into voxel space. LASSO was performed using the `lasso_rocha()` function called by `predict()` a standard function in the CANLab toolbox (<https://github.com/canlab>). This function automatically determines the  $\lambda$  hyper-parameter. The number of retained coefficients (lasso number = 120) was determined by MSE. The model was trained, via LOO-CV, to predict the intensity level (1 – 5). LASSO-PCR is a machine-learning-based-based regression technique. Due to its penalization method, LASSO regression simultaneously performs feature selection and model estimation<sup>27</sup>. This procedure is similar to that which was used to development the neurologic pain signature (NPS)<sup>30</sup> and the picture-induced negative emotion signature (PINES)<sup>29</sup>. This procedure was performed twice. First, for the intent ratings and second for the observer inference ratings. The results were two brain-based models of signal *intent* and socioemotional *inference*, respectively. Model training accuracy was assessed by calculating prediction-outcome correlations (Pearson's  $r$ ) across each CV-fold.

**Bootstrap hypothesis test.** To determine which brain areas made reliable contributions to prediction, we used a bootstrap hypothesis test (5000 samples) over the model weights. For each voxel, we resampled participants with replacement. Each bootstrap replicate was created using MATLAB's `bootstrap` function, which samples the rows of the data matrix with replacement and computes a statistic (here, the mean coefficient) on each bootstrap sample. We generated a distribution of coefficients for each voxel. To conduct a significance test we shifted the bootstrap distribution to zero by subtracting the mean coefficient. Next, we computed a standardized z-score for the original coefficients using the standard deviation of the bootstrap distribution. Finally, we calculated p-values using the normal cumulative distribution function (`normcdf`) to determine statistical significance. We used the  $p$  values to threshold the maps to interpret feature importance.

**Model validation.** To validate the models, we applied it to analogous intent and inference brain activity during the auditory-only and visual-only trials. These validation maps also had five levels of target self-reported emotion intensity and five levels of inferred emotion intensity. Validation trials were "held out," meaning that they were not included in training or tested on during training.

There were two validation sets—one for each model—an intent validation set, and an inference validation set. A prediction-outcome correlation (Pearson's  $r$ ) between the model and each validation image was calculated for each subject ( $N=100$ ). Then we averaged participant's  $r$  values and compared them within and across the validation sets. The validity of the model was established, first, by applying it to its own validation set (i.e., the intent model was tested on the intent validation set) and, next, by testing if the average performance was positively greater than zero in a two-tailed single-sample  $t$ -test.

To further validate the models, we tested its specificity to its own validation set. We did this by applying the model to the opposite validation set (i.e., the intent model was tested on the inference validation set). Then we tested if the model's average performance was higher on its own validation set versus the opposite validation set via a two-tailed paired  $t$ -test. We expected the intent model to have a higher prediction-outcome correlation on the intent validation set than on the inference validation set. Likewise, we expected the inference model to have a higher prediction-outcome correlation on the inference validation set than the intent validation set.

Finally, an additional post hoc internal validation was performed at the level of the training stimuli (Fig. S10). The purpose of this validation was to test whether stimuli with low overall empathic accuracy performance at the group level contributed to the model trainings in a biased way. To interrogate individual videos, we constructed a new test set of participant-level fMRI images for each video (based on intent and inference rating quintiles). We then applied our two models to the new test data and recorded the prediction-outcome correlations for each video. This is analogous to the validation set/double disassociation analyses in Figs. 2B and 3B (see also Fig. S5 for a schematic) described above. High prediction-outcome correlations indicate the model accurately predicted the test set. A *higher* prediction-outcome correlation on a model's sensitivity test (i.e., Inference model on Inference test data) than on a model's specificity test set (i.e., Inference model on Intent test data) would validate the model for that video trial. We also tested if a video's average empathic accuracy performance ( $r$ ) was predictive of the model's accuracy when tested on that video using a Pearson's correlation (Fig. S10C–D). If low correlation videos contributed to the model training in a biased way, then model accuracy should be inflated for low correlation videos and there would be a strong negative correlation between the overall  $r$  of a video (empathic accuracy metric) and the model's accuracy ( $r$ ).

**Characterization of pattern weights.** The unthresholded predictive weight map (see Fig. 1) was input into the NeuroSynth Image Decoder (<https://neurosynth.org/decode/>) to quantitatively compare it to images in the NeuroSynth database. This allows us to assess what behaviors and functions are most associated with our patterns of brain activity across the published literature. The top 50 terms that loaded onto the map were pruned down to remove single brain regions and redundant concepts (like theory of mind and ToM). Then, a word cloud was constructed of the top 20 terms. The networks and functions in the NeuroSynth database which were most similar to the model are depicted in the word cloud (Figs. 2 & 3), and the words are scaled by strength of similarity.

To verify that the intent and inference models are not reducible to general processing of affect, we compared our models' patterns with previously published and publicly available models of affect and emotion. We used cosine similarity, a multivariate method of assessing the similarity between two vectors, to make these comparisons

between our models' pattern weights and eight published models of various components of PINES<sup>29</sup>, NPS<sup>30</sup>, the Vicarious Pain Signature (VPS)<sup>68</sup>, Empathic Care & Empathic Distress<sup>32</sup>, Social Rejection<sup>31</sup>, Galvanic Skin Response (GSR)<sup>69</sup>, and Heart Rate (Heart<sup>69</sup>; see Fig. S7). These models are available freely online (<https://github.com/canlab>). Cosine similarity ranges from -1 to 1, where values closer to 0 mean the two vectors are orthogonal or perpendicular to each other. When the value is closer to 1, it means the angle is smaller and images/brain patterns are more similar. When the value is closer to -1, it means the images are more opposite.

### Comparison of the intent and inference

We tested the separability of participant-level intent and inference coefficient maps at each level of intensity with a linear support vector machine (SVM; C = 1, optimizer = Adam (adaptive moment estimation), LOO-CV) where "inference" maps were the positive class and "intent" maps were the negative class. This classification was performed across each level of intensity, and a receiver operating characteristic (ROC) curve was calculated across all the CV folds (within subject; Fig. 4). In this analysis, a classification accuracy significantly greater than chance indicates that the activity patterns are linearly separable, and therefore, represent unique neural processes<sup>70</sup>. To better understand the brain regions that distinguish between intent and inference, we performed a bootstrap (5000 samples) hypothesis test over the predictive voxel weights for the classifier trained at the highest level of intensity (level 5) and then applied a loose threshold ( $P < 0.05$ ) so that the brain regions where intent and inference maximally diverge could be compared (Fig. 4b). This thresholding was repeated also for the classifiers trained at the other levels of intensity as well for completeness (Fig. S10).

### Model alignment and empathic accuracy

To test how our models relate to an observer's empathic accuracy, we took the dot-product between each model and individual subject's ( $N = 100$ ) maps of empathic accuracy and added in the intercept to calculate the model's rating prediction on the test data. Then we correlated the resulting predictions from low empathic accuracy trials and high empathic accuracy trials across all participants. Low empathic accuracy trials are when there is maximal discordance between target and observer ratings (i.e., target rated 5 and observer rated 1) and high empathic accuracy trials are when there is no difference between target and observer ratings (i.e., target rated 5 and observer rated 5; see "Derivation of empathic accuracy ratings" section above for details). Importantly, these low and high empathic accuracy maps were linearly separable via SVM (see Fig. S12). Finally, the alignment analysis was validated on empathic accuracy maps from the visual-only and auditory-only trials, which are independent from the intent and inference model training data, to ensure that results were not biased. Two-tailed independent z-tests of the correlation difference were used to compare alignment scores ( $r$  values) and estimate confidence intervals.

During first-level modeling, the empathic accuracy regressors may overlap in time with the intent and inference regressors because empathic accuracy is partially derived from target and observer ratings. Therefore, first we confirmed the separability of the intent and inference subject-level beta maps from the empathic accuracy subject-level beta maps before testing how the alignment of the intent and inference models are related to empathic accuracy. The separability of the intent and inference beta maps from the empathic accuracy beta maps at two levels of empathic accuracy (low accuracy and high accuracy) were confirmed using a linear SVM (C = 1, optimizer = Andre, LOO-CV). This classification was performed across each level of intensity vs the low and high empathic accuracy maps. The intent and inference maps were separable from the low and high empathic accuracy maps (accuracy range = 57–67%; SE = 1.5–1.6%;  $P < 0.001$ ).

### Functional connectivity analysis

Raw fMRI timeseries data<sup>71</sup> were cropped into the 24 video stimuli and then preprocessed with CANLabCore by removing the same nuisance regressors described earlier. The Brainetomme Atlas<sup>33</sup> was used to parcel out activity in 273 brain regions (one region in the cerebellum was excluded for missing data so 272 are in the full model). Activity within those regions was averaged spatially and extracted across each video timeseries for each participant. Next, pairwise distance matrices of the ROI timeseries were constructed for each video that appeared in the audiovisual channel and then averaged within each person (see Fig. S17).

Next, we correlated degree centrality of the nodes with overall empathic accuracy (average Pearson's  $r$  across audiovisual stimuli for each participant) across individuals to assess which nodes in the network were related to forming an inference. One participant (Subject 022 that had a technical error in counterbalancing) was excluded from this analysis due to having had an unbalanced number of video data.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

### Data availability

The processed data generated in this study as well as the raw behavioral data have been deposited in the OSF database [<https://osf.io/cs3km/>] (<https://doi.org/10.17605/osf.io/cs3km>). The raw fMRI data are available on OpenNeuro under accession code ds006111 [<https://openneuro.org/datasets/ds006111>] (<https://doi.org/10.18112/openneuro.ds006111.v1.0.0>).

### Code availability

All code for this manuscript is available at <https://osf.io/cs3km/> (<https://doi.org/10.17605/osf.io/cs3km>)<sup>72</sup> and <https://github.com/canlab/CanlabCore>.

### References

1. Holt-Lunstad, J., Smith, T. B. & Layton, J. B. Social relationships and mortality risk: A meta-analytic review. *PLoS Med.* **7**, e1000316 (2010).
2. Cacioppo, J. T. & Cacioppo, S. Older adults reporting social isolation or loneliness show poorer cognitive function 4 years later. *Evid. Based Nurs.* **17**, 59–60 (2014).
3. Fakoya, O. A., McCorry, N. K. & Donnelly, M. Loneliness and social isolation interventions for older adults: A scoping review of reviews. *BMC Public Health* **20**, 129 (2020).
4. Luanaigh, C. O. & Lawlor, B. A. Loneliness and the health of older people. *Int. J. Geriatr. Psychiatry* **23**, 1213–1221 (2008).
5. Goodwin, R., Cook, O. & Yung, Y. Loneliness and life satisfaction among three cultural groups. *Personal. Relatsh.* **8**, 225–230 (2001).
6. Mehu, M. & Scherer, K. R. A psycho-ethological approach to social signal processing. *Cogn. Process.* **13**, 397–414 (2012).
7. Gilbert, K., Hall, K. & Codd, R. T. Radically open dialectical behavior therapy: Social signaling, transdiagnostic utility and current evidence. *Psychol. Res. Behav. Manag.* **13**, 19–28 (2020).
8. Hall, J. A., Andrzejewski, S. A. & Yopchick, J. E. Psychosocial correlates of interpersonal sensitivity: A meta-analysis. *J. Nonverbal Behav.* **33**, 149–180 (2009).
9. Cassidy, S., Ropar, D., Mitchell, P. & Chapman, P. Can adults with autism spectrum disorders infer what happened to someone from their emotional response?: Inferences from reactions. *Autism Res.* **7**, 112–123 (2014).
10. Vucurovic, K., Caillies, S. & Kaladjian, A. Neural correlates of theory of mind and empathy in schizophrenia: An activation likelihood estimation meta-analysis. *J. Psychiatr. Res.* **120**, 163–174 (2020).

11. Button, K. S., Browning, M., Munafò, M. R. & Lewis, G. Social inference and social anxiety: Evidence of a fear-congruent self-referential learning bias. *J. Behav. Ther. Exp. Psychiatry* **43**, 1082–1087 (2012).
12. Baez, S. et al. Theory of mind and its relationship with executive functions and emotion recognition in borderline personality disorder. *J. Neuropsychol.* **9**, 203–218 (2015).
13. Sened, H. et al. Empathic accuracy and relationship satisfaction: a meta-analytic review. *J. Fam. Psychol.* **31**, 742 (2017).
14. Chang, L. J. et al. Endogenous variation in ventromedial prefrontal cortex state dynamics during naturalistic viewing reflects affective experience. *Sci. Adv.* **7**, eabf7129 (2021).
15. Pugh, Z. H., Choo, S., Leshin, J. C., Lindquist, K. A. & Nam, C. S. Emotion depends on context, culture and their interaction: Evidence from effective connectivity. *Soc. Cogn. Affect. Neurosci.* **17**, 206–217 (2022).
16. Skerry, A. E. & Saxe, R. A common neural code for perceived and inferred emotion. *J. Neurosci.* **34**, 15997–16008 (2014).
17. Spunt, R. P. & Adolphs, R. The neuroscience of understanding the emotions of others. *Neurosci. Lett.* **693**, 44–48 (2019).
18. Zaki, J., Weber, J., Bolger, N. & Ochsner, K. The neural bases of empathic accuracy. *Proc. Natl Acad. Sci. USA* **106**, 11382–11387 (2009).
19. Bush, K. A. et al. Brain states that encode perceived emotion are reproducible but their classification accuracy is stimulus-dependent. *Front. Hum. Neurosci.* **12**, 262 (2018).
20. Kragel, P. A. & LaBar, K. S. Multivariate neural biomarkers of emotional states are categorically distinct. *Soc. Cogn. Affect. Neurosci.* **10**, 1437–1448 (2015).
21. Kragel, P. A., Reddan, M. C., LaBar, K. S. & Wager, T. D. Emotion schemas are embedded in the human visual system. *Sci. Adv.* **5**, eaaw4358 (2019).
22. Ong, D. C. et al. Modeling emotion in complex stories: The stanford emotional narratives dataset. *IEEE Trans. Affect. Comput.* **12**, 579–594 (2021).
23. Zaki, J. & Ochsner, K. The need for a cognitive neuroscience of naturalistic social cognition. *Ann. N.Y. Acad. Sci.* **1167**, 16–30 (2009).
24. Eickhoff, S. B., Milham, M. & Vanderwal, T. Towards clinical applications of movie fMRI. *NeuroImage*, **217** (2020).
25. Jääskeläinen, I. P., Sams, M., Glerean, E. & Ahveninen, J. Movies and narratives as naturalistic stimuli in neuroimaging. *NeuroImage* **224**, 117445 (2021).
26. Grimm, P. Social Desirability Bias. In J. Sheth & N. Malhotra (Eds.), *Wiley International Encyclopedia of Marketing* (1st ed.) (2010).
27. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc.: Ser. B (Methodol.)* **58**, 267–288 (1996).
28. Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C. & Wager, T. D. Large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods* **8**, 665–670 (2011).
29. Chang, L. J., Gianaros, P. J., Manuck, S. B., Krishnan, A. & Wager, T. D. A sensitive and specific neural signature for picture-induced negative affect. *PLOS Biol.* **13**, e1002180 (2015).
30. Wager, T. D. et al. An fMRI-based neurologic signature of physical pain. *N. Engl. J. Med.* **368**, 1388–1397 (2013).
31. Woo, C. W. et al. Separate neural representations for physical pain and social rejection. *Nat. Commun.* **17**, 5380 (2014).
32. Ashar, Y. K., Andrews-Hanna, J. R., Dimidjian, S. & Wager, T. D. Empathic care and distress: Predictive brain markers and dissociable brain systems. *Neuron* **94**, 1263–1273.e4 (2017).
33. Fan, L. et al. The human brainnetome atlas: A new brain atlas based on connectional architecture. *Cereb. Cortex* **26**, 3508–3526 (2016).
34. Izard, C. E. Basic emotions, natural kinds, emotion schemas, and a new paradigm. *Perspect. Psychological Sci.* **2**, 260–280 (2007).
35. Scheliga, S. et al. Neural correlates of multisensory integration in the human brain: an ALE meta-analysis. *Rev. Neurosci.* **34**, 223–245 (2023).
36. Barrett, L. F. The theory of constructed emotion: An active inference account of interoception and categorization. *Soc. Cogn. Affect. Neurosci.* **12**, 1–23 (2017).
37. Russell, J. A. Core affect and the psychological construction of emotion. *Psychological Rev.* **110**, 145–172 (2003).
38. Quandt, L. C., Lee, Y.-S. & Chatterjee, A. Neural bases of action abstraction. *Biol. Psychol.* **129**, 314–323 (2017).
39. Aminoff, E. M., Kveraga, K. & Bar, M. The role of the para-hippocampal cortex in cognition. *Trends Cogn. Sci.* **17**, 379–390 (2013).
40. Blakemore, S.-J., Bristow, D., Bird, G., Frith, C. & Ward, J. Somatosensory activations during the observation of touch and a case of vision-touch synesthesia. *Brain* **128**, 1571–1583 (2005).
41. Hooker, C. I., Verosky, S. C., Germine, L. T., Knight, R. T. & D'Esposito, M. Neural activity during social signal perception correlates with self-reported empathy. *Brain Res.* **1308**, 100–113 (2010).
42. Jafari, M. et al. The human primary somatosensory cortex encodes imagined movement in the absence of sensory information. *Commun. Biol.* **3**, 757 (2020).
43. Schaefer, M. et al. Empathy-related brain activity in somatosensory cortex protects from tactile priming effects: A pilot study. *Front. Hum. Neurosci.* **14**, 142 (2020).
44. Genzer, S., Ong, D. C., Zaki, J. & Perry, A. Mu rhythm suppression over sensorimotor regions is associated with greater empathic accuracy. *Soc. Cogn. Affect. Neurosci.* **17**, 788–801 (2022).
45. Jospe, K. et al. The contribution of linguistic and visual cues to physiological synchrony and empathic accuracy. *Cortex* **132**, 296–308 (2020).
46. Jospe, K. et al. Impaired empathic accuracy following damage to the left hemisphere. *Biol. Psychol.* **172**, 108380 (2022).
47. Zhao, Z., Sened, H. & Tamir, D. I. Egocentric projection is a rational strategy for accurate emotion prediction. *J. Exp. Soc. Psychol.* **109**, 104521 (2023).
48. Craig, K. et al. Evaluating metacognitive self-reports: systematic reviews of the value of self-report in metacognitive research. *Metacognition Learn.* **15**, 155–213 (2020).
49. Polimeni, J. R. & Lewis, L. D. Imaging faster neural dynamics with fast fMRI: A need for updated models of the hemodynamic response. *Prog. Neurobiol.* **207**, 102174 (2021).
50. Mauss, I. B. et al. Don't hide your happiness! Positive emotion dissociation, social connectedness, and psychological functioning. *J. Personal. Soc. Psychol.* **100**, 738–748 (2011).
51. Adler, N. E., Epel, E. S., Castellazzo, G. & Ickovics, J. R. Relationship of subjective and objective social status with psychological and physiological functioning: Preliminary data in healthy, White women. *Health Psychol.* **19**, 586–592 (2000).
52. Russell, J. A. A circumplex model of affect. *J. Personal. Soc. Psychol.* **39**, 1161–1178 (1980).
53. Esteban, O. et al. fMRIprep: A robust preprocessing pipeline for functional MRI. *Nat. Methods* **16**, 111–116 (2019).
54. Esteban, O. et al. fMRIprep: A robust preprocessing pipeline for functional MRI (23.0.2) [Computer software]. Zenodo. (2023).
55. Gorgolewski, K. et al. Nipype: A flexible, lightweight and extensible Neuroimaging Data Processing Framework in Python. *Front. Neuroinformatics* **5**, 1662–5196 (2011).
56. Gorgolewski, K. et al. Nipype: A flexible, lightweight and extensible neuroimaging data processing framework in Python. 0.13.1 (0.13.1) [Computer software]. Zenodo. (2017).
57. Tustison, N. J. et al. N4ITK: Improved N3 Bias Correction. *IEEE Trans. Med. Imaging* **29**, 1310–1320 (2010).
58. Dale, A. M., Fischl, B. & Sereno, M. I. Cortical surface-based analysis. *NeuroImage* **9**, 179–194 (1999).

59. Klein, A. et al. Mindboggling morphometry of human brains. *PLOS Computational Biol.* **13**, e1005350 (2017).
60. Fonov, V., Evans, A., McKinstry, R., Almlie, C. & Collins, D. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage* **47**, S102 (2009).
61. Avants, B., Epstein, C., Grossman, M. & Gee, J. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Med. Image Anal.* **12**, 26–41 (2008).
62. Zhang, Y., Brady, M. & Smith, S. Segmentation of brain MR images through a hidden Markov random field model and the expectation–maximization algorithm. *IEEE Trans. Med. Imaging* **20**, 45–57 (2001).
63. Cox, R. W. AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers Biomed. Res.* **29**, 162–173 (1996).
64. Jenkinson, M. Fast, automated, N-dimensional phase-unwrapping algorithm. *Magn. Reson. Med.* **49**, 193–197 (2003).
65. Greve, D. N. & Fischl, B. Accurate and robust brain image alignment using boundary-based registration. *NeuroImage* **48**, 63–72 (2009).
66. Behzadi, Y., Restom, K., Liao, J. & Liu, T. T. A component-based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage* **37**, 90–101 (2007).
67. Power, J. D. et al. Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage* **84**, 320–341 (2014).
68. Krishnan, A. et al. Somatic and vicarious pain are represented by dissociable multivariate brain patterns. *eLife* **5**, e15166 (2016).
69. Eisenbarth, H., Chang, L. J. & Wager, T. D. Multivariate brain prediction of heart rate and skin conductance responses to social threat. *J. Neurosci.* **36**, 11987–11998 (2016).
70. Reddan, M. C., Wager, T. D. & Schiller, D. Attenuating neural threat expression with imagination. *Neuron* **100**, 994–1005.e4 (2018).
71. Reddan, M. C., Ong, D., Wager, T. D. & Zaki, J. Stanford emotional narratives fMRI dataset. *OpenNeuro* <https://doi.org/10.18112/openneuro.ds006111.v1.0.0> (2025).
72. Reddan, M. C. Code and data for neural signatures of emotional inference and experience align during social consensus. *OSF*. [Dataset] <https://doi.org/10.17605/osf.io/cs3km> (2023).

## Acknowledgements

Research reported in this publication was supported National Institutes of Health (NIH) Award Number R01MH112560 (J.Z.) and the National Center for Advancing Translational Sciences of the NIH under Award Number K12TR004411 (M.C.R.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

## Author contributions

J.Z. and T.D.W. designed research; D.C.O., S.M., and I.K. performed research; M.C.R. analyzed data; M.C.R. and J.Z. wrote the paper.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-59931-8>.

**Correspondence** and requests for materials should be addressed to Marianne C. Reddan or Jamil Zaki.

**Peer review information** *Nature Communications* thanks Yina Ma, Samuel Nastase and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025