Behavioral/Cognitive

Hostile Attribution Bias Shapes Neural Synchrony in the Left Ventromedial Prefrontal Cortex during Ambiguous Social Narratives

Yizhou Lyu (吕奕洲),[1]* Zishan Su (苏紫杉),[1]* Dawn Neumann,[2] Kimberly L. Meidenbauer,[3] and Yuan Chang Leong (梁元彰)[1,4]

[1]Department of Psychology, University of Chicago, Chicago 60637, Illinois, [2]Department of Physical Medicine and Rehabilitation, Indiana University School of Medicine, Indianapolis 46202, Indiana, [3]Department of Psychology, Washington State University, Pullman 99164, Washington, and [4]Neuroscience Institute, The University of Chicago, Chicago 60637, Illinois

Hostile attribution bias refers to the tendency to interpret social situations as intentionally hostile. While previous research has focused on its developmental origins and behavioral consequences, the underlying neural mechanisms remain underexplored. Here, we employed functional near-infrared spectroscopy (fNIRS) to investigate the neural correlates of hostile attribution bias. While undergoing fNIRS, male and female participants listened to and provided attribution ratings for 21 hypothetical scenarios where a character's actions resulted in a negative outcome for the listener. Ratings of hostile intentions were averaged to measure hostile attribution bias. Using intersubject representational similarity analysis, we found that participants with similar levels of hostile attribution bias exhibited higher levels of neural synchrony during narrative listening, suggesting shared interpretations of the scenarios. This effect was localized to the left ventromedial prefrontal cortex (VMPFC) and was particularly prominent in scenarios where the character's intentions were highly ambiguous. We then grouped participants into high and low bias groups based on a median split of their hostile attribution bias scores. A similarity-based classifier trained on the neural data classified participants as having high or low bias with 75% accuracy, indicating that the neural time courses during narrative listening was systematically different between the two groups. Furthermore, hostile attribution bias correlated negatively with attributional complexity, a measure of one's tendency to consider multifaceted causes when explaining behavior. Our study sheds light on the neural mechanisms underlying hostile attribution bias and highlights the potential of using fNIRS to develop nonintrusive and cost-effective neural markers of this sociocognitive bias.

Key words: fNIRS; hostile attribution bias; neural synchrony; social cognition; VMPFC

Significance Statement

Inferring the intentions from behavior is crucial for adaptive social functioning. A predisposition toward interpreting intentions as hostile is a significant predictor of interpersonal conflict and aggressive tendencies. Using functional near-infrared spectroscopy, we found that individual differences in hostile attribution bias shaped neural synchrony in the ventromedial prefrontal cortex while processing real-world social situations. Additionally, we were able to distinguish between participants with high and low hostile attribution bias from their neural activity. These results reveal how subjective interpretations of social situations are influenced by hostile attribution bias and reflected in the temporal dynamics of brain activity. Our findings lay the groundwork for future studies aimed at understanding the neurobiological basis of sociocognitive biases and interventions that mitigate these biases.

Received July 6, 2023; revised Dec. 20, 2023; accepted Jan. 7, 2024.

Author contributions: Y.L., K.L.M., and Y.C.L. designed research; Y.L., Z.S., and Y.C.L. performed research; D.N. and K.L.M. contributed unpublished reagents/analytic tools; Y.L., Z.S., and Y.C.L. analyzed data; Y.L., Z.S., and Y.C.L. wrote the paper.

We thank the University of Chicago fNIRS user group, John Veillette, Susan Levine, and members of the Motivation and Cognition Neuroscience Lab for helpful comments on the study. We also thank the University of Chicago Neuroscience Institute Shared Equipment Award for providing the fNIRS device.

*Y.L and Z.S. contributed equally to this work.

The authors declare no competing financial interests.

Correspondence should be addressed to Yuan Chang Leong at ycleong@uchicago.edu.

https://doi.org/10.1523/JNEUROSCI.1252-23.2024

Introduction

Picture two strangers walking down a crowded sidewalk, brushing shoulders as they pass. While one of them might view this as an innocuous and unavoidable outcome, the other might perceive it as a deliberate hostile act. Hostile attribution bias refers to a predisposition to perceive others' actions as hostile and is thought to reflect a skewed system of appraisals and expectancies that biases social judgments (Epps and Kendall, 1995; Dodge, 2006; Klein Tuente et al., 2019). The tendency to infer hostile intentions predisposes an individual to respond aggressively,

resulting in a self-fulfilling prophecy where perceived hostility begets actual hostility. Indeed, hostile attribution bias has been tightly linked to increased physical and relational aggression, impaired social relationships, and poor mental health (Pettit et al., 2010; Dodge et al., 2015; Smith et al., 2016). Studying the neural basis of hostile attribution bias can inform targeted interventions that reduce aggressive behavior and foster healthier relationships.

While the implications of hostile attribution bias on maladaptive behavior are extensively documented, its neurobiological underpinnings remain relatively underexplored. Prior studies have used structural MRI and lesion-mapping methods to identify brain regions associated with individual differences in hostile attribution bias and aggression (Grafman et al., 1996; Yang and Raine, 2009; Cristofori et al., 2016; Quan et al., 2019). A consistent finding emerging from these studies is that structural differences (e.g., morphological variation or lesions) in the prefrontal cortex (PFC) are associated with the tendency to make hostile attributions and behave aggressively. For example, Quan et al. (2019) found that larger gray matter volume in the left orbitofrontal cortex (OFC) was associated with higher trait-level hostile attribution bias. Furthermore, gray matter volume in the left OFC mediated the effects of hostile attribution bias on participants' willingness to endorse violence. In line with these findings, noninvasive stimulation of the PFC modulates aggressive tendencies (Hortensius et al., 2012; Dambacher et al., 2015; Choy et al., 2018).

The prior work suggests a pivotal role of the PFC in the manifestation and regulation of hostile attribution bias. These studies have examined the structural correlates of hostile attribution bias and its downstream behavioral effects, but it remains unclear if and how prefrontal regions are engaged during the ongoing processing of social information that ultimately gives rise to hostile attributions. Furthermore, there is a lack of spatial specificity in the prefrontal regions involved, with some studies highlighting ventromedial regions (Grafman et al., 1996; Quan et al., 2019) and others highlighting dorsolateral regions (Cristofori et al., 2016; Choy et al., 2018). Additionally, existing functional neuroimaging studies have focused on aggression and violence (Yang and Raine, 2009; Fanning et al., 2017) and not the attributional biases that result in the hostile interpretation of social cues.

The goal of this study is to investigate the dynamic engagement of prefrontal regions during real-time processing of social situations and how they contribute to hostile attribution bias. We used functional near-infrared spectroscopy (fNIRS) to measure the activity in the PFC as participants listened to scenarios created to measure hostile attribution bias (Epps and Kendall, 1995). The cost-effectiveness of fNIRS allowed for a larger sample size, which is particularly important for individual difference analyses. The scenarios described social situations where a character acted in a manner that led to a negative outcome for the listener. Hostile attribution bias was measured by evaluating participants' hostility ratings of the characters' actions. Due to mixed evidence implicating different prefrontal regions, we measured activity across the PFC. We then used intersubject representational similarity analysis (IS-RSA; Finn et al., 2020) to test the hypothesis that individuals with similar levels of hostile attribution bias would have similar neural dynamics while listening to the narratives, reflecting similar interpretations of the scenarios. Thus, our paradigm allowed us to relate activity of specific prefrontal regions to individual differences in the tendency to attribute hostile intentions, paving the way for a deeper understanding of the neurobiological mechanisms underpinning hostile attribution bias.

Materials and Methods

Participants

Sixty-four individuals participated in the study. Experimental procedures were approved by the University of Chicago Institutional Review Board, and all participants provided informed consent prior to the start of the study. Participants received either 1 course credit or $15 for the hour-long experiment. All participants self-reported having native proficiency in English, no hearing or speech comprehension disorders, no serious head injury, and no neurological/psychiatric disorders and were not taking psychiatric medication. Individuals were recruited from the University of Chicago community through the research participation system managed by the Department of Psychology (SONA systems). The study advertisement and consent form indicated that the study was an investigation of the neural basis of how people understand narratives. Four participants had incomplete data due to technical difficulties with the recording device and were excluded from analyses. Additionally, data from two participants were excluded because of unusable data (see fNIRS data acquisition and preprocessing), yielding an effective sample size of 58 participants (20 males, 37 females, 1 nonbinary; Table 1). All participants who completed the study reported that they were able to comprehend the narratives.

Experimental procedures

Participants were scanned using fNIRS while they listened to 21 narrated scenarios (total duration, 13 min 30 s; average duration, 38.57 s; Fig. 1*A*). These scenarios were taken from a scenario-based questionnaire used to measure hostile attribution bias in adult samples (Epps and Kendall, 1995). In each scenario, the character in the story acted in a way that resulted in a hypothetical negative outcome for the listener (e.g., a former employer forgetting to submit a letter of recommendation). Prior work conducted by the author of these scenarios shows that, on average, they elicit hostile, benign, and ambiguous attributions of the character's intentions (hostile, benign, and ambiguous scenarios; seven in each category). We used this questionnaire over others due to the relatively large number of scenarios and because the scenarios tended to be longer and described nuanced social situations. We recorded a research assistant narrating each scenario in a neutral tone. The use of audio rather than written stimuli allowed for more precise temporal control over how participants processed the scenarios.

While listening, participants were asked to imagine and react to the scenarios as if they were happening to them. Following each scenario, participants rated how well they understood the recording (comprehension ratings), how angry they would be if the incident happened to them (anger ratings), if they thought the character intended their actions (intentionality ratings), if they thought the person's actions were hostile (hostility ratings), and the extent to which they blamed the character for their actions (blameworthiness ratings). All ratings were collected from a 0 to 9 point scale and rescaled to a 1–9 point scale to facilitate comparison

Table 1. Participant demographics

Variables	$N = 58$
Age, mean (SD)	21.9 (5.5)
Gender, *n* (%)	
Male	20 (34.5%)
Female	37 (63.8%)
Nonbinary	1 (1.7%)
Race, *n* (%)	
White/European American	22 (37.9%)
Asian/Asian American or Pacific Islander	22 (37.9%)
Hispanic/Latino	4 (6.9%)
Black/African American	3 (5.2%)
Middle Eastern/Arab American	1 (1.7%)
Other	6 (10.3%)

with prior studies that have used these scenarios (Epps and Kendall, 1995; Neumann et al., 2015, 2021).

There was a 5 s rest period between the end of the last rating and the start screen of the next scenario. Participants were then instructed to indicate with a key press when they were ready to listen to the next scenario. This was followed by a 5 s "loading screen" before the next recording started. On average, 38.2 s (SE = 4.95 s) elapsed between the end of the previous recording and the beginning of the next. The scenarios were presented in a fixed order across participants (Extended Data Fig. 1-1). We followed the same presentation order used when these scenarios are administered in clinical settings (Epps and Kendall, 1995; Neumann et al., 2015, 2021). Presenting stimuli in a fixed order is also a common methodological choice in fMRI studies examining interindividual variability in neural synchrony (Chen et al., 2020; Finn et al., 2020; Baek et al., 2023). In particular, a fixed order minimizes intersubject variability in neural synchrony due to presentation order and increases sensitivity to detect intersubject variability that arises from endogenous differences between participants.

One potential concern is that there would be carry-over emotional responses from one recording to the next. As the primary goal of the current study was to test if interindividual differences in hostile attribution bias are associated with differences in neural responses, rather than to contrast neural responses across conditions, the advantages of a fixed order were judged to outweigh the potential cost. We minimized carry-over effects by imposing 5 s intervals before and after each scenario. In addition, the task proceeded to the next scenario only when participants indicated with a button press that they were ready to listen to the next recording.

fNIRS data acquisition and preprocessing

fNIRS data were collected using a NIRSport2 fNIRS unit (NIRx Medical Technologies) with a layout of 20 channels composed of eight light sources and seven detectors. Spatial positions were standardized across participants using the unambiguously illustrated (UI) 10/10 external positioning system (Jurcak et al., 2007). Data were collected at a sampling rate of 10.1725 Hz at wavelengths of 760 and 850 nm. Due to a limited number of available optodes, it was not possible to obtain whole-brain fNIRS coverage. Optodes were placed to optimize coverage of the PFC due to extensive prior work implicating the region in aggression (Harmon-Jones and Sigelman, 2001; Yang and Raine, 2009; Cristofori et al., 2016; Choy et al., 2018), social attributions (Forbes and Grafman, 2010; Spunt and Lieberman, 2012; Wagner et al., 2019; Elder et al., 2023), and the subjective interpretation of social narratives (Yeshurun et al., 2017; Finn et al., 2018; Nguyen et al., 2019; Leong et al., 2020; Dieffenbach et al., 2021), suggesting that the region may play a critical role in the manifestation of hostile attribution bias.

Data preprocessing was performed using custom MATLAB scripts that utilized the Homer2 analysis package (Huppert et al., 2009). Channels were identified as unusable and removed if detector saturation occurred for longer than 2 s or if the signal's power spectrum exceeded a quartile coefficient of dispersion of 0.29 over the course of the scan (Dieffenbach et al., 2021). Data from two participants were discarded for having >50% unusable channels. For each channel, raw light intensity values were converted to optical density by taking the negative logarithm of the ratio between light intensity and the mean light intensity within each channel. A bandpass filter with a frequency range of 0.005–0.15 Hz was then applied to reduce the influence of low-frequency drift, respiration, and cardiac activity. Motion artifacts were identified and corrected using targeted principal components analysis (Yücel et al., 2014). Specifically, motion artifacts were identified as optical density values that exceeded five standard deviations or an absolute change in signal amplitude of two within a 1 s interval. A principal components analysis was then performed to remove 80% of the variance in the 1 s around the identified artifact.

Optical density values were then converted to changes in oxygenated (HbO), deoxygenated (HbR), and total (HbT) hemoglobin concentrations using the modified Beer–Lambert law with a ppf value of 6. All analyses utilized the HbO signal, which has stronger signal amplitude, higher correlation to fMRI BOLD signals, and better signal-to-noise ratio than HbR and HbT (Strangman et al., 2002; Tong and Frederick, 2010; Duan et al., 2012). Finally, the HbO time course of each participant was shifted by 4.5 s to account for the lag in the hemodynamic response (Huppert et al., 2006) and *z*-scored across time.

Intersubject representational similarity analyses

We employed IS-RSA (Nguyen et al., 2019; Chen et al., 2020; Finn et al., 2020; Fig. 2*A*) to examine the relationship between hostile attribution bias and neural synchrony during narrative listening. IS-RSA builds on classic RSA (Kriegeskorte et al., 2008), where the similarity between different experimental conditions is compared with the similarity in neural responses to the experimental conditions. A correlation between the two would indicate a correspondence between how internal task representations are hypothesized to differ and the neural representation in a given brain region. Here, instead of comparing responses between experimental conditions within a participant, we compared the responses to the same experimental stimuli across participants to test whether participants with similar levels of hostile attribution bias exhibit greater similarity in their neural responses.

For each participant, we first computed their hostile attribution bias score as the average hostility rating across the 21 stories. We then constructed a *behavioral distance matrix* by calculating the absolute difference in hostile attribution scores between every pair of participants. As such, the behavioral distance matrix reflects the similarity structure in hostile attribution across participants, with smaller distances denoting pairs of participants with similar levels of hostile attribution bias. Our fNIRS analyses relied on the method of intersubject correlation (ISC) and diverged from both block and event-related designs that are more commonly used in fNIRS studies. Specifically, ISC analyses focus on the temporal fluctuations of brain activity during stimulus presentation, rather than on changes from baseline in a given block or in response to a specific event (Nastase et al., 2019). For each fNIRS channel, we extracted and concatenated the activity time course while the participant listened to the 21 scenarios. Data collected while participants were completing the subjective ratings and the 5 s rest period were not included in the analyses. The concatenated time courses were then *z*-scored across time.

While ISC studies usually calculate *neural similarity* across individuals, we computed *neural distance* (i.e., the inverse of neural similarity) as is typical for representational similarity analyses (Kriegeskorte et al., 2008). This also allowed us to test different distance metrics to assess the robustness of our results. For each channel, we constructed a *neural distance matrix* by computing the correlation distance (i.e., 1 - Pearson's *r*) of the concatenated time courses between every pair of participants. The neural distance matrix reflects the similarity structure in the temporal dynamics of neural activity between pairs of participants at that corresponding channel, with smaller distances denoting pairs of participants whose activity time course was more similar. As the distance matrices were symmetrical along the diagonal, only the lower triangle of each matrix (excluding the diagonal) was retained.

We then computed the Pearson's correlation between the behavioral distance matrix and the neural distance matrix for each channel. A positive correlation indicates a channel where activity time courses were more synchronous between participants who were similar in hostile attribution bias. We computed IS-RSA with Pearson's correlation instead of Spearman's correlation because the former is sensitive to differences in magnitude between measurements, and not just their rank order. This is particularly relevant in our study as we were interested in examining IS-RSA separately across hostile, ambiguous, and benign scenarios. The magnitude of neural distances between pairs is likely to differ across scenario types, even if the rank order does not. To test the robustness of our results, we also ran the IS-RSA with Spearman's correlation.

Statistical significance was assessed using a Mantel test (Mantel, 1967; Kriegeskorte et al., 2008). Specifically, we recomputed the IS-RSA with behavioral distance matrices constructed from data where the identities of the participants were shuffled such that the distance matrix no longer reflects the similarity structure observed in the data. This procedure was repeated 10,000 times to generate a null distribution, with the *p* value calculated as the proportion of the null distribution with an *r* value that was more positive than the empirical *r* value. We then corrected for multiple

comparisons across 20 channels by controlling for the false discovery rate (FDR; Benjamini and Hochberg, 1995) of $q < 0.05$ using the *p.adjust* function implemented in R. An alternative method for computing neural distance is to compute the Euclidean distance between activity time courses. As we did not have theoretical reasons to prefer correlation distance over Euclidean distance, we reran the IS-RSA using neural distance matrices constructed from the Euclidean distance between the activity time courses of participant pairs. This allowed us to test if our results were robust to different metrics of neural distance.

To examine the relationship between hostile attribution bias and neural synchrony for hostile, ambiguous, and benign scenarios, we reran the IS-RSA for each narrative type. In all three analyses, the behavioral distance matrix was constructed from the average hostility ratings across all 21 scenarios (i.e., the participant's hostile attribution bias scores) and not just the ratings from scenarios in the narrative type. Our rationale was that hostile attribution bias was a trait measure and we wanted to be consistent in the measure that we used rather than have it vary based on the scenarios participants were listening to. Taking the mean hostility ratings across all scenarios provided the best estimate of an individual's level of hostile attribution bias, which we then related to temporal patterns of neural activity in different situations. Separate neural distance matrices were computed from the neural time courses during hostile, ambiguous, and benign scenarios.

To examine the relationship between neural synchrony and intersubject similarity in intentionality attributions, anger, and assessments of blameworthiness, we reran the IS-RSA with behavioral distance matrices constructed from intentionality, anger, and blameworthiness ratings. For each behavioral measure, we first computed the average rating across the 21 scenarios for each participant. The behavioral distance matrix was then computed from the absolute difference in average ratings of the corresponding measure between each pair of participants and correlated with the neural distance matrix.

Additionally, we tested if an *Anna Karenina* (AnnaK) model (Finn et al., 2020; Baek et al., 2023) of intersubject representational similarity would account for the neural data. The *Anna Karenina* model is named after the opening line of the Tolstoy novel, "All happy families are alike; each unhappy family is unhappy in its own way," and captures the hypothesis that some participants might be alike, while others are idiosyncratic. We constructed an *AnnaK* behavioral distance matrix where each cell is computed as the average hostile attribution bias score of each pair of participants. In the AnnaK behavioral matrix, pairs of participants with low hostile attribution bias would have a low distance value, while pairs of participants with high hostile attribution bias would have a high distance value. We then computed the correlation between the AnnaK behavioral matrix and the neural distance matrix of each channel. A positive correlation would indicate that neural responses were more synchronized among participants with low hostile attribution bias but were more idiosyncratic among participants with high hostile attribution bias. In contrast, a negative correlation would suggest the opposite; that is, neural responses were more synchronized among participants with high hostile attribution bias but were more idiosyncratic among participants with low hostile attribution bias. Statistical significance was again assessed using Mantel tests and corrected at FDR $q < 0.05$.

The IS-RSAs measure the association between pairwise dissimilarity in hostile attribution bias and pairwise dissimilarity in neural activity. The resulting r values were thus not a direct measure of the correlation between individual differences in hostile attribution bias and neural activity. To estimate the strength of the underlying relationship between hostile attribution bias and neural activity, we ran a simulation study to map RSA r values (i.e., the correlation of distance matrices) to the direct correlation between the two underlying variables. We adopted a grid-search approach of correlation values ranging from 0 to 0.5 (interval = 0.02). For each correlation value, we generated two random vectors with the corresponding correlation. Each vector contained 58 observations to match the sample size of the study. We computed the pairwise distance matrix within each vector and correlated the two distance matrices to obtain the RSA r value. This procedure was repeated 1,000 times for each correlation value and the mean RSA r value was computed. A best-fit line of the mean RSA r values across correlation values was estimated using LOESS fitting.

Visualization of IS-RSA results using multidimensional scaling

We projected the neural distance matrix onto a two-dimensional space using multidimensional scaling (MDS) as implemented in the MATLAB mdscale function with default settings. Individual participants were then colored according to their hostile attribution bias score. To visualize the relationship between neural distance and hostile attribution bias, we applied a Gaussian smoothing kernel with a bandwidth of 0.2 across participants' hostile attribution bias scores on a grid spanning from −0.8 to 0.8 on both axes with a resolution of 100 by 100 points. The resulting smoothed values were translated to a color spectrum ranging from blue (representing lower hostile attribution scores) to red (representing higher hostile attribution scores). The MDS plot provides an interpretable visual summary of the neural distances in relation to individual differences in hostile attribution bias.

Behavioral analyses across narrative types

Hostility ratings were averaged for each participant across scenarios separately for hostile, ambiguous, and benign scenarios. Paired-sample t tests were used to test whether mean hostility ratings differed between narrative types. The analysis was repeated for intentionality, anger, and blameworthiness ratings. To examine if intersubject variability in hostility ratings differed between narrative types, we computed the standard deviation in hostility ratings for each scenario. Two-sample t tests were used to test whether the average standard deviation of hostility ratings differed between narrative types.

Analyses on mean fNIRS activity across narrative types

We computed the mean neural activity in the left ventromedial prefrontal cortex (VMPFC; channel 11) across the duration for each scenario. We then averaged the mean neural activity across scenarios separately for hostile, ambiguous, and benign narratives. Paired-sample t tests were then used to test if mean neural activity at each channel was different between (1) hostile and ambiguous narratives, (2) hostile and benign narratives, and (3) ambiguous and benign narratives.

Synchrony-based classification analysis

We adapted a synchrony-based classification approach used in prior fNIRS and fMRI studies (Yeshurun et al., 2017; Dieffenbach et al., 2021) to assess the extent to which we were able to identify participants with high and low hostile attribution bias. Participants were first grouped into high and low hostile attribution groups using a median split of their hostile attribution bias scores ($n = 29$ in each group). For each iteration of a leave-one-out cross-validation procedure, a participant was held out as the test data. For the remaining participants, we averaged the neural time courses of each channel separately for each group to obtain a "template" time course that captured the temporal dynamics of neural activity for the respective group. The activity time course for the held-out participant was then correlated with each of the two template time courses. The participant was classified as high hostile attribution bias if the correlation was higher with the high hostile attribution bias template and classified as low hostile attribution bias if the correlation was higher with the low hostile attribution bias template. Classification accuracy was computed separately at each channel. Statistical significance was assessed using a nonparametric permutation test by comparing the true classification accuracy to a null distribution generated by re-running the analysis 10,000 times with shuffled group labels. Results were then thresholded at an FDR q of <0.05.

We also ran the classification analysis with a more extreme grouping of participants. Low bias participants were defined as participants with hostile attribution bias scores in the lowest third of the sample ($n = 18$; score range, 1.71–3.30), while high bias participants were defined as participants with scores in the top third ($n = 18$; score range, 4.41–6.17). The groups had 18 rather than 19 participants due to ties in scores near the group boundaries.

Spatial localization and visualization of fNIRS channels
To spatially localize and visualize the fNIRS results, approximate MNI coordinates of each fNIRS channel were determined using an anchor-based probabilistic conversion atlas (Tsuzuki et al., 2012). The fNIRS data were then converted to NifTI files using xjView (https://www.alivelearn.net/xjview) and projected onto the cortical surface using BrainNet Viewer (http://www.nitrc.org/projects/bnv/; Xia et al., 2013).

Pre-experimental surveys
Prior to the start of the experiment, participants completed the following behavioral questionnaires:

Aggression Questionnaire (Buss and Perry, 1992). The Aggression Questionnaire (AQ) is a self-assessment questionnaire that measures overall aggression, with subscales for physical aggression, verbal aggression, anger, and hostility. Participants were presented with 29 statements and asked to rate on a 5-point scale the extent to which the statement is characteristic of themselves. Participants' total score on the AQ was used for analyses. The AQ is widely used as a measure of trait aggression (Archer, 2004; Sekine et al., 2008; da Cunha-Bang et al., 2017).

Revised Social Connectedness Scale (Lee et al., 2001). The revised Social Connectedness Scale (SCS-R) is a 20-item questionnaire that is widely used to measure social connection (Sandstrom and Dunn, 2014; Aknin et al., 2022). Participants were asked to rate the extent to which they agree or disagree with the 20 items on a 6-point scale. Participant's total score on the SCS-R was used for analyses.

Attributional Complexity Scale (Fletcher et al., 1986). The Attributional Complexity Scale (ACS) is a 28-item questionnaire used to measure attributional complexity (i.e., an individual's tendency and ability to consider complex and multiple causes for their own and others' behavior). The scale has been externally validated with studies demonstrating that individuals who score highly on attributional complexity spontaneously produced more causes when explaining behavior and selected more complex causal attributions for behavioral events (Fletcher et al., 1986). Attributional complexity has also been shown to be positively correlated with measures of empathy and perspective-taking and negatively correlated with depression and anxiety (Joireman, 2004; Fast et al., 2008). Participant's total score on the ACS was used for analyses.

Code and data accessibility
Data and analysis code can be accessed at https://github.com/lyulouisa/Hostile_Attribution_Bias.

Results

Fifty-eight participants were scanned using fNIRS as they listened and provided ratings to 21 hypothetical scenarios (total duration, 13 min 30 s; Fig. 1*A*). Hostility ratings were correlated with ratings of intentionality (average $r = 0.82$; SE = 0.018; $t_{(58)} = 46.5$; $p < 0.001$), blameworthiness (average $r = 0.60$; SE = 0.026; $t_{(58)} = 23.0$; $p < 0.001$), and anger across scenarios (average $r = 0.41$; SE = 0.029; $t_{(58)} = 14.32$; $p < 0.001$), suggesting that participants were more likely to attribute intentionality and blame to characters they rated as hostile, and would also be more angry at those characters. Across participants, average hostility ratings were correlated with average intentionality ($r = 0.91$; $t_{(58)} = 16.7$; $p < 0.001$), blameworthiness ($r = 0.76$; $t_{(58)} = 8.99$; $p < 0.001$), and anger ratings ($r = 0.81$; $t_{(58)} = 10.51$; $p < 0.001$), indicating that the participants who were more likely to perceive the characters as acting in a hostile manner were the same participants who were likely to attribute blame and intentionality to the characters, as well as experience greater anger in these situations. For each participant, we computed the average hostility ratings across all scenarios as an individual difference measure of hostile attribution bias (hostile attribution bias score; Fig. 1*B*).

Hostile attribution bias shapes neural synchrony in the VMPFC

We employed IS-RSA (Nguyen et al., 2019; Finn et al., 2020; Fig. 2*A*) to identify neural correlates of hostile attribution bias. The intuition behind the analytical approach is that participants

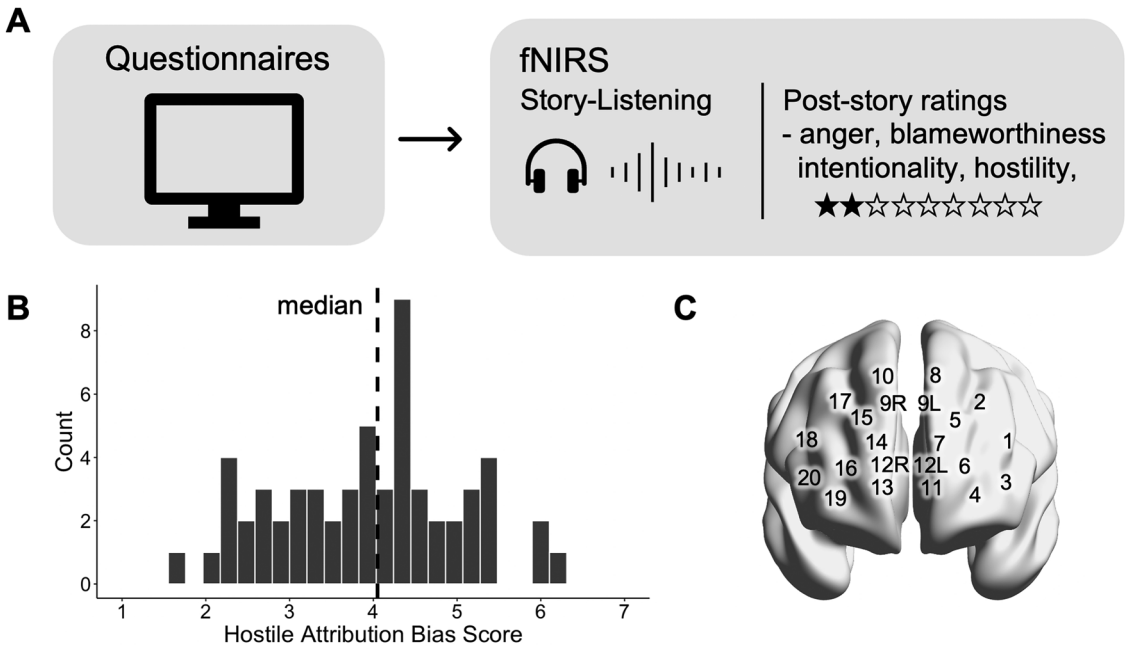

Figure 1. Experimental design. ***A***, Prior to the experiment, participants completed a series of behavioral questionnaires that measured individual differences in aggression, social connectedness, and attributional complexity. They then listened to 21 narrated social scenarios while undergoing fNIRS. At the end of each story, participants provided hostility, intentionality, blameworthiness, and anger ratings. See Extended Data Table 1-1 for title, narrative category, and duration of 21 scenarios. ***B***, We computed each participant's hostile attribution bias score as the average hostility ratings across all scenarios. Dotted line indicates sample median. ***C***, Spatial location of the 20 fNIRS channels projected onto the cortical surface.

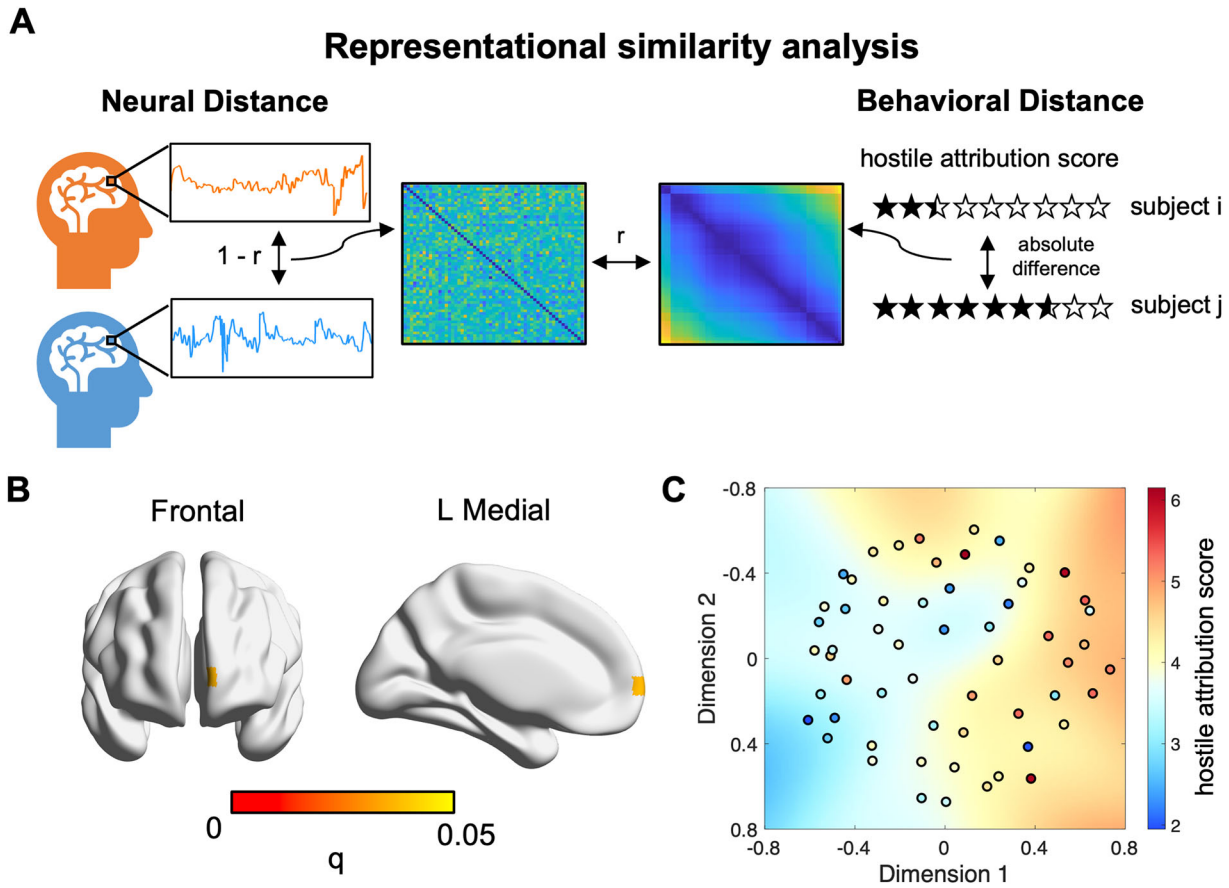

Figure 2. Hostile attribution bias shapes neural synchrony in the VMPFC. ***A***, Schematic of IS-RSA. For each channel, we constructed a neural distance matrix by calculating the 1 - Pearson's correlation between the time courses of every pair of participants. We then correlated each neural distance matrix with a behavioral distance matrix constructed from the absolute difference in hostile attribution bias scores for every pair of participants. Extended Data Figure 2-1 shows correspondence between IS-RSA *r* and correlation between underlying variables. ***B***, Left VMPFC (channel 11) activity time course was more similar in individuals with similar levels of hostile attribution bias. Extended Data Figure 2-2 reports results from all channels. Extended Data Figure 2-3 reports results from IS-RSA with Euclidean distance. Extended Data Figure 2-4 reports results from IS-RSA with intentionality, anger, and blameworthiness ratings. ***C***, Neural distance between participants at the left VMPFC projected onto two dimensions using MDS. Each point represents a participant and the distance between points represents the neural distance between corresponding participants. The color gradient reflects hostile attribution score, generated by applying a Gaussian smoothing kernel across the plot.

with similar levels of hostile attribution bias would process the narratives in a similar fashion and brain regions associated with hostile attribution bias would exhibit similar temporal dynamics while listening to the narratives. We used IS-RSA to assess the similarity between pairwise distances in hostile attribution bias and pairwise distances in neural time courses. The resulting IS-RSA *r* value reflects the extent to which similarity in hostile attribution bias tracked similarity in neural responses to the narratives.

After controlling for multiple comparisons, the IS-RSA yielded a single significant channel corresponding to the left VMPFC (channel 11, VMPFC; $r = 0.131$; $p < 0.001$; $q = 0.006$; Fig. 2*B*; Extended Data Fig. 2-2). A simulation study indicated that an IS-RSA *r* value of 0.13 corresponded approximately to a correlation of $r = 0.38$ between individual differences in hostile attribution bias and neural activity (see Materials and Methods; Extended Data Fig. 2-1). Pairwise similarity in hostile attribution bias did not track neural similarity in the right VMPFC (channel 13; $r = 0.011$; $p = 0.330$; $q = 0.770$). The same pattern of results was observed when the similarity between neural and behavioral matrices were computed using Spearman's correlation (left VMPFC, channel 11: $r = 0.119$, $p < 0.001$, $q = 0.006$; right VMPFC, channel 13: $r = 0.014$, $p = 0.276$, $q = 0.276$).

To visualize the IS-RSA results, we applied MDS to the left VMPFC neural distance matrix. MDS represents distances between pairs of observations in a lower-dimensional space, where observations that are more similar to one another are placed closer together. The resulting MDS plot indicated that participants with similar hostile attribution biases were indeed closer in neural distance (Fig. 2*C*). Specifically, participants with stronger hostile attribution biases were more likely to occupy the upper right triangle of the plot while those with weaker biases were more likely to occupy the lower left triangle.

As a test of whether our results were robust to a different metric of neural distance, we repeated the IS-RSA with neural distance matrices constructed using the Euclidean distance in channel time course between pairs of participants. This yielded identical results, where the channel corresponding to the left VMPFC was the only channel where similarity in levels of hostile attribution bias correlated with similarity in neural responses ($r = 0.131$; $p < 0.001$; $q = 0.006$; Extended Data Fig. 2-3). The VMPFC has been previously shown to encode subjective evaluations and appraisals (Hutcherson et al., 2015; Chang et al., 2021). Here, our results extend these earlier findings by providing evidence that the subjective interpretation of social situations is shaped by hostile attribution bias and encoded in temporal dynamics of VMPFC activity.

In a series of exploratory analyses, we examined the relationship between neural synchrony and intersubject similarity in

intentionality attributions, anger, and assessments of blameworthiness. We reran the IS-RSA with behavioral distance matrices constructed from intentionality, anger, and blameworthiness ratings. Similar results were observed with intentionality ratings, in that pairwise distances in intentionality ratings between participants were correlated with intersubject dissimilarity in left VMPFC activity time courses ($r = 0.118$; $p < 0.001$; $q = 0.006$; Extended Data Fig. 2-4). In contrast, no significant channels were observed with anger and blameworthiness ratings (all qs > 0.10; Extended Data Fig. 2-4). These results suggest that neural synchrony in the left VMPFC was shaped by shared interpretations of hostile intentions, rather than by shared negative affect due to the negative actions of the characters.

We also tested whether our data would be explained by the *Anna Karenina* (*AnnaK*) model of neural similarity, where hostile attribution bias modulated the level of idiosyncrasy in participants' neural responses (Finn et al., 2020; Baek et al., 2023; see Materials and Methods). A positive AnnaK correlation would indicate that neural responses were similar between participants with low levels of hostile attribution bias and idiosyncratic for participants with high levels of hostile attribution bias. In contrast, a negative AnnaK correlation would indicate the opposite. We found no significant correlations across the 20 channels (all qs > 0.30), indicating that the AnnaK models did not provide a good account of the relationship between hostile attribution bias and neural responses.

Divergence in VMPFC synchrony was driven by narratives with ambiguous intent

The 21 audio narratives can be categorized into scenarios that on average elicit hostile, ambiguous, or benign attributions of the character's intentions (Epps and Kendall, 1995). In our sample, benign scenarios ($M = 2.97$; SE = 0.09) received lower hostility ratings than hostile scenarios ($M = 4.52$; SE = 0.12; mean difference = −1.54; SE = 0.14; $t_{(57)} = -10.9$; $p < 0.001$) and ambiguous scenarios ($M = 4.34$; SE = 0.12; mean difference = −1.37; SE = 0.12; $t_{(57)} = -11.0$; $p < 0.001$). Hostility ratings were not significantly different between hostile and ambiguous scenarios (mean difference = −0.17; SE = 0.11; $t_{(57)} = -1.54$; $p = 0.13$; Fig. 3*A*). The same pattern was observed for ratings of anger, blameworthiness, and intentionality (Extended Data Fig. 3-1).

To investigate if hostile attribution bias modulated how these narratives were processed, we ran the IS-RSA separately for hostile, benign, and ambiguous narratives. When the analysis was applied to data from the ambiguous narratives, we replicated the IS-RSA results where the left VMPFC (channel 11) was the only channel for which the neural distance correlated significantly with pairwise distances in hostile attribution bias scores ($r = 0.142$; $p < 0.001$; $q = 0.002$). In contrast, the IS-RSA yielded no significant channels for hostile and benign narratives (all qs > 0.10). In other words, neural activity diverged between participants with different levels of hostile attribution bias when listening to narratives where the intent of the social other was ambiguous, but not when listening to narratives where the intent was clearly hostile or benign.

One interpretation of these results is that narratives where social intentions were clearly hostile or benign left less room for divergent interpretations, such that participants were likely to make similar attributions regardless of individual differences in hostile attribution bias. However, when intentions were ambiguous, participants' interpretations were more strongly shaped by their intrinsic predispositions to attribute hostile or benign intentions, thus neural responses diverged between participants with varying levels of hostile attribution bias. This interpretation is consistent with several studies which have suggested that it is in ambiguous situations, where situational cues are lacking, that hostile attribution bias exerts the strongest influence (Combs et al., 2007; Pinkham et al., 2014; Neumann et al., 2020, 2021). Consequently, ambiguous scenarios may provide sufficient information for evaluating hostile attribution bias.

To test this possibility in our behavioral data, we computed the standard deviation (SD) in hostility ratings for each scenario and assessed how the average SD differed between narrative types. Average SD was indeed higher for ambiguous scenarios (mean $SD_{ambiguous} = 1.94$; SE = 0.08) than that for benign scenarios (mean $SD_{benign} = 1.73$; SE = 0.09). Unexpectedly, the average SD was lower for ambiguous scenarios than that for hostile scenarios (mean $SD_{hostile} = 2.10$; SE = 0.09), suggesting hostile scenarios elicited more intersubject variability in hostility ratings. One possible explanation could be that some participants were unwilling to report overly negative hostile attributions even for clearly hostile situations, resulting in a larger spread of ratings across participants for the hostile scenarios. This would also be consistent with the finding where there were no significant differences in mean hostility ratings between hostile and ambiguous scenarios. We note that none of the comparisons in intersubject variability were statistically significant ($p > 0.10$), which was expected since there were only seven scenarios in each narrative type. Thus, we advise caution when interpreting these behavioral results.

No significant differences in mean left VMPFC activity between narrative types

The IS-RSA analyses indicated that temporal dynamics in left VMPFC activity were modulated by hostile attribution bias during ambiguous scenarios but not in hostile and benign scenarios. We next examined whether mean left VMPFC activity over the course of a scenario differed between the three narrative types. For each participant, we calculated the mean left VMPFC activity over the course of each scenario and averaged the mean activity separately for hostile, ambiguous, and benign narratives. Across participants, left VMPFC activity was not significantly different between the three types of narratives (hostile > ambiguous: mean difference = 0.03, SE = 0.03 $t_{(56)} = 0.949$, $p = 0.347$; hostile > benign: mean difference = 0.05, SE = 0.03, $t_{(56)} = 1.388$, $p = 0.171$; ambiguous > benign: mean difference = 0.01, SE = 0.04, $t_{(56)} = 0.328$, $p = 0.744$).

We note that our paradigm was not optimized for examining mean activity across conditions. In particular, averaging over the 40-second-long scenarios may have obscured transient neural responses that occurred at specific moments during the scenarios. In contrast, the ISC analyses allowed us to sidestep this limitation by focusing specifically on temporal fluctuations in the activity time courses rather than mean differences in overall activity.

Synchrony-based classification of hostile attribution bias

Our results indicate that neural activity diverged between individuals with varying levels of hostile attribution bias. Could we then identify participants with high or low levels of hostile attribution bias based on their neural responses to the narratives? We adapted a synchrony-based classification approach used in prior fNIRS and fMRI studies (Fig. 4*A*; Yeshurun et al., 2017; Dieffenbach et al., 2021) to test this possibility. We grouped participants into those with high and low levels of hostile attribution bias based on a median split of their hostile attribution scores. Following a leave-one-out cross-validation approach, we computed the similarity between each participant's neural time

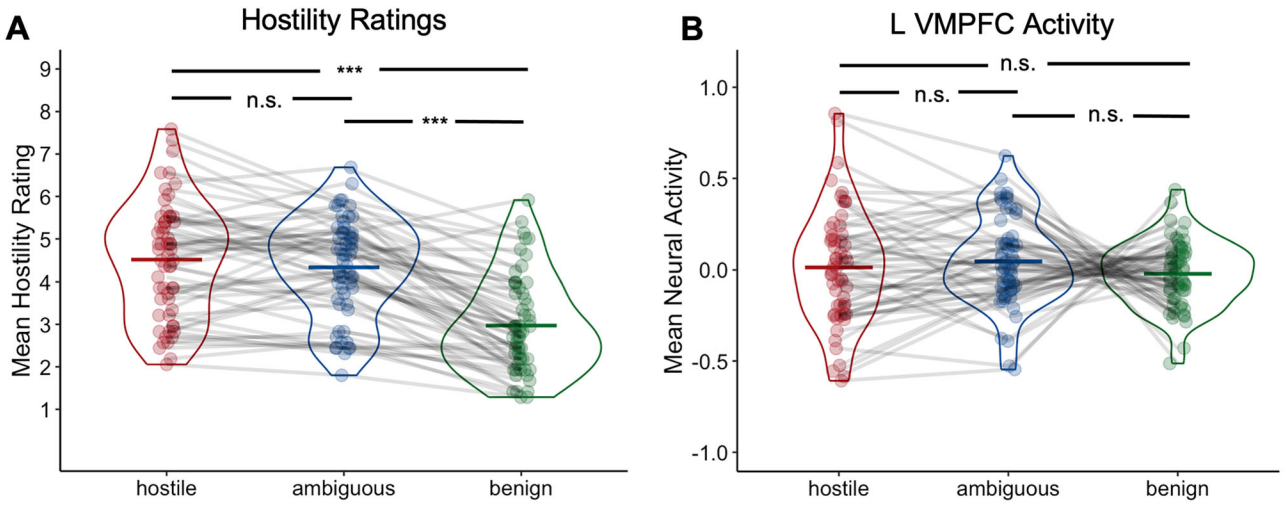

Figure 3. Violin plots comparing hostile, ambiguous, and benign scenarios. ***A***, Mean hostility rating for each narrative type. Extended Data Figure 3-1 reports differences in mean intentionality, anger, and blameworthiness ratings. ***B***, Mean neural activity in the left VMPFC (channel 11) during each narrative type. Each data point indicates an individual participant. Thin gray lines connect data from the same participant. Thick colored lines indicate corresponding group mean. ***$p < 0.001$; n.s., not significantly different.

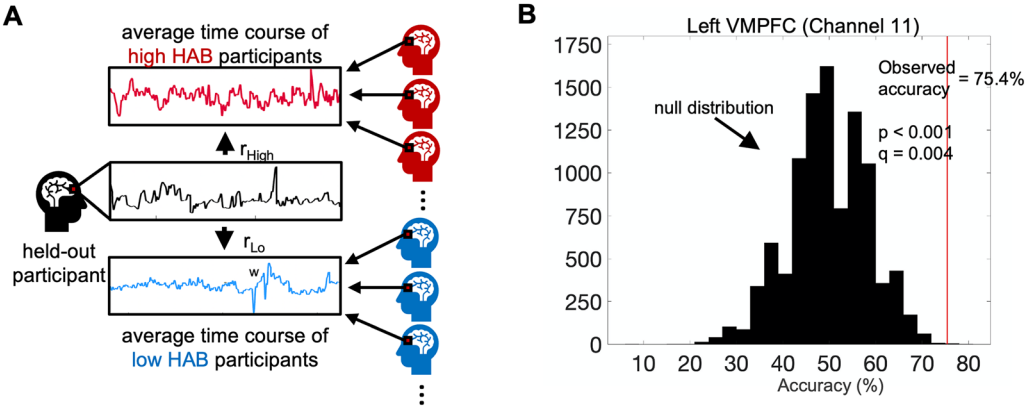

Figure 4. Synchrony-based classification of hostile attribution bias. ***A***, Schematic of classification approach. Following a leave-one-out cross-validation approach, we correlated the activity time course of a held-out participant with the average time course of participants with high and low hostile attribution bias. The held-out participant was classified into the group whose time course they were more correlated to. ***B***, Classification accuracy in the left VMPFC (channel 11). Red line indicates observed classification accuracy. Histogram shows null distribution generated from a nonparametric permutation test. Extended Data Figure 4-1 reports results of all channels. Extended Data Figure 4-2 reports results when participants in the middle tertile were excluded from the analyses.

course and the average neural time course of each group. These average neural time courses can be thought of as a "template" time course reflecting how the typical high and low hostile attribution bias participant processed the 21 narratives. If there was a stronger correlation with the high hostile attribution bias template, we classified the participant as having high hostile attribution bias; and if there was a stronger correlation with the low hostile attribution bias template, we classified the participants as having low hostile attribution bias.

Classification accuracy was significantly above chance only in the left VMPFC (channel 11: accuracy = 75.4%, $p < 0.001$, $q = 0.004$; Fig. 4*B*; Extended Data Fig. 4-1), indicating that time courses were sufficiently consistent within a group, and distinct between groups, such that we could distinguish between participants with high and low bias based on their neural time courses alone. Grouping participants based on a median split might have incorrectly grouped individuals with scores close to the median as high or low bias when they did not belong to either group. In an exploratory analysis, we reran the classification analysis with more stringent criteria for grouping participants. We defined high bias participants as participants with the highest tertile of hostile attribution scores ($n = 18$) and low bias participants as participants with the lowest tertile of hostile attribution scores ($n = 18$). Consistent with the earlier results, classification accuracy was significantly above chance only in the left VMPFC (accuracy = 86.1%; $p < 0.001$; $q = 0.004$; Extended Data Figure 4-2). Notably, accuracy was >10% higher, suggesting that discarding participants in the middle of the scale led to a more consistent neural signal in each group.

The classification results are consistent with the lack of an AnnaK relationship, as they indicate neither high nor low bias participants were particularly idiosyncratic. Instead, participants with high hostile attribution bias were similar to other participants with high hostile attribution bias, and participants with low hostile attribution bias were similar to other participants with low hostile attribution bias. The results also suggest that interpretations of the narratives tended to be consistent within each group, such that each scenario had a "canonical" hostile

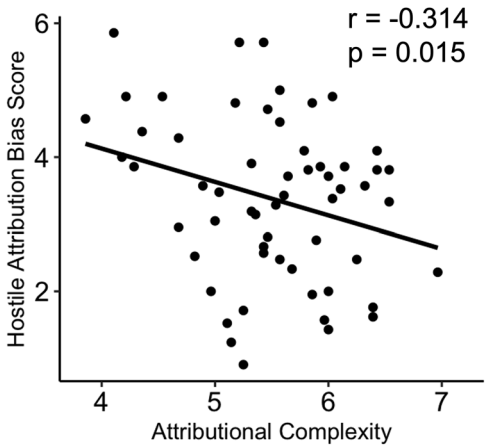

Figure 5. Attributional complexity is negatively correlated with hostile attribution bias. Attributional complexity refers to one's tendency and ability to consider complex and multiple causes for their own and others' behavior. Each data point indicates an individual participant.

interpretation shared by participants with high hostile attribution bias, and a "canonical" benign interpretation shared by participants with low hostile attribution bias.

Attributional complexity is negatively correlated with hostile attribution bias

Understanding the predictors of hostile attribution bias can reveal the underlying psychological factors and might allow researchers to develop interventions that mitigate the bias. Here, we consider two potential predictors: attributional complexity and social connectedness. Attributional complexity was measured using the ACS (Fletcher et al., 1986) and refers to one's tendency and ability to consider complex and multiple causes for their own and others' behavior, taking into account both external (i.e., situational) and internal (i.e., dispositional) factors. On the one hand, attributional complexity might exacerbate hostile attribution bias as the individual might be more likely to generate complex explanations to rationalize their hostile attributions. On the other hand, attributional complexity might buffer against hostile attribution bias by encouraging a more nuanced understanding of others' behavior. Here, we found that attributional complexity was negatively correlated with hostile attribution bias ($r = -0.314$; $t_{(58)} = -2.52$; $p = 0.015$; Fig. 5), suggesting that individuals who tend to infer complex attributions were less likely to make hostile attributions.

Social connectedness, as measured using the SCS-R (Lee et al., 2001), was not correlated with hostile attribution bias ($r = -0.120$; $t_{(58)} = -0.95$; $p = 0.346$). As aggression is often cited as a downstream consequence of hostile attributions (Neumann et al., 2015; Coccaro et al., 2016; Klein Tuente et al., 2019), we sought to test the relationship between aggression and hostile attribution bias. In our sample, trait aggression, as measured using the AQ (Buss and Perry, 1992), was only marginally correlated with hostile attribution bias ($r = 0.238$; $t_{(58)} = 1.87$; $p = 0.067$), possibly because our sample consisted of healthy participants in a college community while prior studies have often recruited special populations (e.g., patients with clinical diagnosis, incarcerated individuals).

Discussion

We combined fNIRS and a narrative listening paradigm to examine how neural activity during the processing of social situations is influenced by individual differences in hostile attribution bias. During narrative listening, temporal dynamics of left VMPFC activity were more synchronous between individuals with similar levels of hostile attribution bias. The effect of hostile attribution bias on neural synchrony was only observed during narratives where the intention of the social other was ambiguous, suggesting that uncertain social contexts were more susceptible to the influence of intrinsic attributional biases. Left VMPFC time courses between participants with high and low hostile attribution bias were systematically different such that a synchrony-based classifier distinguished between the two groups from neural activity alone. We also found that people who infer more complex attributions of others' behaviors were less likely to make hostile attributions. Together, our results suggest that hostile attribution bias influences subjective interpretations of social situations via differential responses in the left VMPFC.

The VMPFC has been implicated in a broad range of complex cognitive functions, including evaluating the subjective value of economic and moral decisions (Bartra et al., 2013; Hutcherson et al., 2015), inferring traits and intentions from behavior (Young et al., 2010; Leopold et al., 2012; Kestemont et al., 2016), and narrative comprehension (van Kesteren et al., 2010; Burin et al., 2014; Yeshurun et al., 2017). One proposed account that unifies these disparate findings is that the VMPFC integrates information about the external world with internal states and prior beliefs to generate "affective meaning"—the subjective appraisal of a situation, experience, or object, including its relevance to oneself and the affective states it engenders (Roy et al., 2012; Chang et al., 2021). In line with this proposal, a recent fMRI study by Chang et al. (2021) mapped patterns of activity in the VMPFC onto specific affective states while participants watched a 45 min television episode. The temporal dynamics of these patterns were largely idiosyncratic across individuals, consistent with the hypothesis that the VMPFC encodes subjective affective meaning that varies from person to person. Our results suggest that temporal dynamics in the VMPFC track with the subjective interpretation of ambiguous social situations and are in line with these earlier findings. Furthermore, our work extends past work by demonstrating that VMPFC dynamics are not necessarily idiosyncratic but can be synchronized between individuals with similar sociocognitive biases. Specifically, our results suggest that hostile attribution bias acts as an intrinsic "prior" that shapes subjective interpretations and VMPFC time courses when processing social information.

We found that hostile attribution bias was associated with neural synchrony in the left, but not the right, VMPFC. Previous studies have documented a similar left–right asymmetry in the PFC, with regions in the left hemisphere showing stronger associations with anger, aggression, and hostility than corresponding regions in the right hemisphere (Gansler et al., 2009; Harmon-Jones et al., 2010; Wang et al., 2018). For example, larger gray matter volume in the left OFC, a region immediately adjacent to the left VMPFC, was associated with higher levels of hostile attribution bias (Quan et al., 2019). Left frontal activity, as measured using EEG, was higher in participants who had just been insulted or socially rejected, with the magnitude of left frontal response correlating with subsequent aggression (Harmon-Jones and Sigelman, 2001; Verona et al., 2009). Furthermore, stimulating the left frontal cortex using noninvasive brain stimulation increased aggression in a laboratory task, while stimulating the right frontal cortex had no effect (Hortensius et al., 2012). It is worthwhile to note, however, that intersubject similarity in anger and blameworthiness ratings did not predict left VMPFC

synchrony in our study. Thus, left VMPFC synchrony did not reflect shared anger at the character in the scenario. Instead, we believe left VMPFC activity was specifically encoding shared interpretations of the character's hostile intentions. Consistent with this hypothesis, similarity in intentionality ratings was also associated with left VMPFC synchrony.

A meta-analysis of 43 structural and functional neuroimaging studies found that reduced structure and function in the lateral prefrontal cortices (LPFCs) was associated with antisocial behavior (e.g., aggression, violence; Yang and Raine, 2009). Furthermore, increased LPFC activity following negative social interactions is associated with decreases in retaliatory aggression (Achterberg et al., 2016, 2020), and disrupting LPFC function using noninvasive brain stimulation prior to negative interactions increases retaliatory aggression (Perach-Barzilay et al., 2013; Choy et al., 2018). In our study, we did not find that the LPFC was differentially engaged depending on participants' level of hostile attribution bias. One possible explanation is that the LPFC is involved in inhibiting aggressive tendencies following negative social interactions, consistent with its role in behavioral inhibition more broadly (Shackman et al., 2009; Aron et al., 2014). In the earlier studies, participants had the opportunity to retaliate against the individual who caused them harm, and the LPFC is engaged to inhibit aggressive behavior. Participants in our task, however, were not able to respond to the hypothetical characters in the scenarios. Instead, the task primed them to consider the intentions of the characters. Without a behavioral response to inhibit, the LPFC is less likely to be engaged, and consequently, the time course of LPFC activity was not modulated by individual differences in hostile attribution bias.

Given the extensive research implicating the PFC in subjective appraisals, social cognition, and aggression, our study focused specifically on the PFC to better understand its role in the manifestation of hostile attribution bias. Our results highlight the role of the VMPFC in encoding subjective interpretations of social situations that were biased by individual differences in hostile attribution bias. There are likely other brain regions involved in this process, including the amygdala, a region often associated with threat detection and processing (Öhman, 2005; Coccaro et al., 2007), which also has dense reciprocal connections with the VMPFC (Barbas, 2000; Kim et al., 2011). Unfortunately, fNIRS is limited to measuring cortical activity near the surface of the brain and is unable to reach deep brain structures such as the amygdala. A whole-brain functional neuroimaging method (e.g., fMRI) will be needed to examine if the amygdala–VMPFC interactions contribute to hostile attribution bias. Another candidate region of interest is the temporoparietal junction (TPJ), which is widely thought to be involved in reasoning about others' intentions and beliefs (Saxe and Kanwisher, 2003; Van Overwalle, 2009). Unlike the amygdala, the TPJ is located closer to the cortical surface (Jiang et al., 2015), and future studies can use fNIRS to examine the relationship between TPJ activity and hostile attribution bias.

In our study, participants who scored higher on attributional complexity exhibited less hostile attribution bias in their behavioral responses to the narratives, a relationship which to our knowledge has not been previously shown. This result suggests that individuals with a propensity to consider multifaceted explanations for behavior are less inclined to interpret ambiguous behaviors as malicious in intent. If so, fostering attributional complexity could be a potential strategy to mitigate hostile attribution bias and ultimately promote healthier social interactions. In line with this hypothesis, a recent study with individuals with traumatic brain injury found that a group-based intervention that promotes thinking about alternative causes of behavior and perspective-taking in simulated and experienced social situations was effective at reducing attributions of intent, blame, anger, and aggression responses to the same 21 scenarios used in the current study (Neumann et al., 2023). Future work could combine the intervention with neural measures to determine its neural basis, which would inform efforts at developing neuroscience-informed markers of hostile attribution bias and intervention efficacy.

Social interaction involves the decoding and interpretation of subtle cues, the recognition of intent, and the formation of appropriate responses based on those interpretations. Individuals with hostile attribution bias have a distorted perception of social intentions, often resulting in unnecessary misunderstandings and conflicts. Here, our results highlight the influence of hostile attribution bias on shaping the neural responses to social situations and demonstrate the viability of using fNIRS to investigate the interplay between neural processes and cognitive biases in social perception. fNIRS offers a unique combination of noninvasiveness, ease of use, and sufficient spatial resolution to localize brain responses to specific regions of the cortex (Yücel et al., 2017). The relative cost-effectiveness makes it an ideal tool for large-scale studies that can more precisely examine individual differences in sociocognitive biases, as well as the development of scalable tools for use in clinical settings. fNIRS thus offers a practical alternative for studying populations with a higher incidence of hostile attribution bias but less amenable to fMRI, including children (Nasby et al., 1980; Dodge et al., 2015) and patients with traumatic brain injury (Neumann et al., 2020, 2021). While our study did not take full advantage of fNIRS's portability and tolerance to motion, the results lay the groundwork for studying hostile attribution bias in more ecologically valid settings (e.g., an unconstrained conversation). Altogether, our study thus paves the way for future research employing fNIRS to better understand the dynamics of social cognition, including how they can break down in the presence of maladaptive biases.

References

Achterberg M, van Duijvenvoorde ACK, Bakermans-Kranenburg MJ, Crone EA (2016) Control your anger! The neural basis of aggression regulation in response to negative social feedback. Soc Cogn Affect Neurosci 11:712–720.

Achterberg M, van Duijvenvoorde ACK, van IJzendoorn MH, Bakermans-Kranenburg MJ, Crone EA (2020) Longitudinal changes in DLPFC activation during childhood are related to decreased aggression following social rejection. Proc Natl Acad Sci U S A 117:8602–8610.

Aknin LB, et al. (2022) Mental health during the first year of the COVID-19 pandemic: a review and recommendations for moving forward. Perspect Psychol Sci 17:915–936.

Archer J (2004) Sex differences in aggression in real-world settings: a meta-analytic review. Rev Gen Psychol 8:291–322.

Aron AR, Robbins TW, Poldrack RA (2014) Inhibition and the right inferior frontal cortex: one decade on. Trends Cogn Sci 18:177–185.

Baek EC, Hyon R, López K, Du M, Porter MA, Parkinson C (2023) Lonely individuals process the world in idiosyncratic ways. Psychol Sci 34:683–695.

Barbas H (2000) Connections underlying the synthesis of cognition, memory, and emotion in primate prefrontal cortices. Brain Res Bull 52:319–330.

Bartra O, McGuire JT, Kable JW (2013) The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. NeuroImage 76:412–427.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Methodol 57:289–300.

Burin DI, Acion L, Kurczek J, Duff MC, Tranel D, Jorge RE (2014) The role of ventromedial prefrontal cortex in text comprehension inferences: semantic coherence or socio-emotional perspective? Brain Lang 129:58–64.
Buss AH, Perry M (1992) The Aggression Questionnaire. J Pers Soc Psychol 63:452–459.
Chang LJ, Jolly E, Cheong JH, Rapuano KM, Greenstein N, Chen P-HA, Manning JR (2021) Endogenous variation in ventromedial prefrontal cortex state dynamics during naturalistic viewing reflects affective experience. Sci Adv 7:eabf7129.
Chen P-HA, Jolly E, Cheong JH, Chang LJ (2020) Intersubject representational similarity analysis reveals individual variations in affective experience when watching erotic movies. NeuroImage 216:116851.
Choy O, Raine A, Hamilton RH (2018) Stimulation of the prefrontal cortex reduces intentions to commit aggression: a randomized, double-blind, placebo-controlled, stratified, parallel-group trial. J Neurosci 38: 6505–6512.
Coccaro EF, McCloskey MS, Fitzgerald DA, Phan KL (2007) Amygdala and orbitofrontal reactivity to social threat in individuals with impulsive aggression. Biol Psychiatry 62:168–178.
Coccaro EF, Fanning JR, Keedy SK, Lee RJ (2016) Social cognition in intermittent explosive disorder and aggression. J Psychiatr Res 83: 140–150.
Combs DR, Penn DL, Wicher M, Waldheter E (2007) The Ambiguous Intentions Hostility Questionnaire (AIHQ): a new measure for evaluating hostile social-cognitive biases in paranoia. Cogn Neuropsychiatry 12: 128–143.
Cristofori I, Zhong W, Mandoske V, Chau A, Krueger F, Strenziok M, Grafman J (2016) Brain regions influencing implicit violent attitudes: a lesion-mapping study. J Neurosci 36:2757–2768.
da Cunha-Bang S, et al. (2017) Violent offenders respond to provocations with high amygdala and striatal reactivity. Soc Cogn Affect Neurosci 12:802–810.
Dambacher F, Schuhmann T, Lobbestael J, Arntz A, Brugman S, Sack AT (2015) Reducing proactive aggression through non-invasive brain stimulation. Soc Cogn Affect Neurosci 10:1303–1309.
Dieffenbach MC, Gillespie GSR, Burns SM, McCulloh IA, Ames DL, Dagher MM, Falk EB, Lieberman MD (2021) Neural reference groups: a synchrony-based classification approach for predicting attitudes using fNIRS. Soc Cogn Affect Neurosci 16:117–128.
Dodge KA (2006) Translational science in action: hostile attributional style and the development of aggressive behavior problems. Dev Psychopathol 18:791–814.
Dodge KA, et al. (2015) Hostile attributional bias and aggressive behavior in global context. Proc Natl Acad Sci U S A 112:9310–9315.
Duan L, Zhang Y-J, Zhu C-Z (2012) Quantitative comparison of resting-state functional connectivity derived from fNIRS and fMRI: a simultaneous recording study. NeuroImage 60:2008–2018.
Elder JJ, Davis TH, Hughes BL (2023) A fluid self-concept: how the brain maintains coherence and positivity across an interconnected self-concept while incorporating social feedback. J Neurosci 43:4110–4128.
Epps J, Kendall PC (1995) Hostile attributional bias in adults. Cogn Ther Res 19:159–178.
Fanning JR, Keedy S, Berman ME, Lee R, Coccaro EF (2017) Neural correlates of aggressive behavior in real time: a review of fMRI studies of laboratory reactive aggression. Curr Behav Neurosci Rep 4:138–150.
Fast LA, Reimer HM, Funder DC (2008) The social behavior and reputation of the attributionally complex. J Res Pers 42:208–222.
Finn ES, Corlett PR, Chen G, Bandettini PA, Constable RT (2018) Trait paranoia shapes inter-subject synchrony in brain activity during an ambiguous social narrative. Nat Commun 9:2043.
Finn ES, Glerean E, Khojandi AY, Nielson D, Molfese PJ, Handwerker DA, Bandettini PA (2020) Idiosynchrony: from shared responses to individual differences during naturalistic neuroimaging. NeuroImage 215: 116828.
Fletcher GJO, Danilovics P, Fernandez G, Peterson D, Reeder GD (1986) Attributional complexity: an individual differences measure. J Pers Soc Psychol 51:875–884.
Forbes CE, Grafman J (2010) The role of the human prefrontal cortex in social cognition and moral judgment. Annu Rev Neurosci 33:299–324.
Gansler DA, McLaughlin NCR, Iguchi L, Jerram M, Moore DW, Bhadelia R, Fulwiler C (2009) A multivariate approach to aggression and the orbital frontal cortex in psychiatric patients. Psychiatry Res Neuroimaging 171: 145–154.
Grafman J, Schwab K, Warden D, Pridgen A, Brown HR, Salazar AM (1996) Frontal lobe injuries, violence, and aggression: a report of the Vietnam head injury study. Neurology 46:1231–1231.
Harmon-Jones E, Sigelman J (2001) State anger and prefrontal brain activity: evidence that insult-related relative left-prefrontal activation is associated with experienced anger and aggression. J Pers Soc Psychol 80:797–803.
Harmon-Jones E, Gable PA, Peterson CK (2010) The role of asymmetric frontal cortical activity in emotion-related phenomena: a review and update. Biol Psychol 84:451–462.
Hortensius R, Schutter DJLG, Harmon-Jones E (2012) When anger leads to aggression: induction of relative left frontal cortical activity with transcranial direct current stimulation increases the anger–aggression relationship. Soc Cogn Affect Neurosci 7:342–347.
Huppert TJ, Hoge RD, Diamond SG, Franceschini MA, Boas DA (2006) A temporal comparison of BOLD, ASL, and NIRS hemodynamic responses to motor stimuli in adult humans. NeuroImage 29:368–382.
Huppert TJ, Diamond SG, Franceschini MA, Boas DA (2009) HomER: a review of time-series analysis methods for near-infrared spectroscopy of the brain. Appl Opt 48:D280–D298.
Hutcherson CA, Montaser-Kouhsari L, Woodward J, Rangel A (2015) Emotional and utilitarian appraisals of moral dilemmas are encoded in separate areas and integrated in ventromedial prefrontal cortex. J Neurosci 35:12593–12605.
Jiang J, Chen C, Dai B, Shi G, Ding G, Liu L, Lu C (2015) Leader emergence through interpersonal neural synchronization. Proc Natl Acad Sci U S A 112:4274–4279.
Joireman J (2004) Relationships between attributional complexity and empathy. J Individ Differ 2:197–202.
Jurcak V, Tsuzuki D, Dan I (2007) 10/20, 10/10, and 10/5 systems revisited: their validity as relative head-surface-based positioning systems. NeuroImage 34:1600–1611.
Kestemont J, Van Mieghem A, Beeckmans K, Van Overwalle F, Vandekerckhove M (2016) Social attributions in patients with ventromedial prefrontal hypoperfusion. Soc Cogn Affect Neurosci 11:652–662.
Kim MJ, Loucks RA, Palmer AL, Brown AC, Solomon KM, Marchante AN, Whalen PJ (2011) The structural and functional connectivity of the amygdala: from normal emotion to pathological anxiety. Behav Brain Res 223: 403–410.
Klein Tuente S, Bogaerts S, Veling W (2019) Hostile attribution bias and aggression in adults - a systematic review. Aggress Violent Behav 46:66–81.
Kriegeskorte N, Mur M, Bandettini PA (2008) Representational similarity analysis - connecting the branches of systems neuroscience. Front Syst Neurosci 2:4.
Lee RM, Draper M, Lee S (2001) Social connectedness, dysfunctional interpersonal behaviors, and psychological distress: testing a mediator model. J Couns Psychol 48:310–318.
Leong YC, Chen J, Willer R, Zaki J (2020) Conservative and liberal attitudes drive polarized neural responses to political content. Proc Natl Acad Sci U S A 117:27731–27739.
Leopold A, Krueger F, dal Monte O, Pardini M, Pulaski SJ, Solomon J, Grafman J (2012) Damage to the left ventromedial prefrontal cortex impacts affective theory of mind. Soc Cogn Affect Neurosci 7:871–880.
Mantel N (1967) The detection of disease clustering and a generalized regression approach. Cancer Res 27:209–220.
Nasby W, Hayden B, DePaulo BM (1980) Attributional bias among aggressive boys to interpret unambiguous social stimuli as displays of hostility. J Abnorm Psychol 89:459–468.
Nastase SA, Gazzola V, Hasson U, Keysers C (2019) Measuring shared responses across subjects using intersubject correlation. bioRxiv:600114.
Neumann D, Malec JF, Hammond FM (2015) The association of negative attributions with irritation and anger after brain injury. Rehabil Psychol 60:155–161.
Neumann D, Sander AM, Perkins SM, Bhamidipalli SS, Witwer N, Combs D, Hammond FM (2020) Assessing negative attributions after brain injury with the ambiguous intentions hostility questionnaire. J Head Trauma Rehabil 35:E450.
Neumann D, Sander AM, Witwer N, Jang JH, Bhamidipalli SS, Hammond FM (2021) Evaluating negative attributions in persons with brain injury: a comparison of 2 measures. J Head Trauma Rehabil 36:E170–E177.
Neumann D, Backhaus SB, Jang J, Bhamadipalli S, Winegardner J, Helton B, Hammond F (2023) Intervention to change attributions that are negative: a feasibility study on reducing anger after brain injury. J Emot Psychopathol 1:72–89.

Nguyen M, Vanderwal T, Hasson U (2019) Shared understanding of narratives is correlated with shared neural responses. NeuroImage 184: 161–170.

Öhman A (2005) The role of the amygdala in human fear: automatic detection of threat. Psychoneuroendocrinology 30:953–958.

Perach-Barzilay N, Tauber A, Klein E, Chistyakov A, Ne'eman R, Shamay-Tsoory SG (2013) Asymmetry in the dorsolateral prefrontal cortex and aggressive behavior: a continuous theta-burst magnetic stimulation study. Soc Neurosci 8:178–188.

Pettit GS, Lansford JE, Malone PS, Dodge KA, Bates JE (2010) Domain specificity in relationship history, social-information processing, and violent behavior in early adulthood. J Pers Soc Psychol 98:190–200.

Pinkham AE, Penn DL, Green MF, Buck B, Healey K, Harvey PD (2014) The social cognition psychometric evaluation study: results of the expert survey and RAND panel. Schizophr Bull 40:813–823.

Quan F, Zhu W, Dong Y, Qiu J, Gong X, Xiao M, Zheng Y, Zhao Y, Chen X, Xia L-X (2019) Brain structure links trait hostile attribution bias and attitudes toward violence. Neuropsychologia 125:42–50.

Roy M, Shohamy D, Wager TD (2012) Ventromedial prefrontal-subcortical systems and the generation of affective meaning. Trends Cogn Sci 16: 147–156.

Sandstrom GM, Dunn EW (2014) Social interactions and well-being: the surprising power of weak ties. Pers Soc Psychol Bull 40:910–922.

Saxe R, Kanwisher N (2003) People thinking about thinking people: the role of the temporo-parietal junction in "theory of mind". NeuroImage 19: 1835–1842.

Sekine Y, et al. (2008) Methamphetamine causes microglial activation in the brains of human abusers. J Neurosci 28:5756–5761.

Shackman AJ, McMenamin BW, Maxwell JS, Greischar LL, Davidson RJ (2009) Right dorsolateral prefrontal cortical activity and behavioral inhibition. Psychol Sci 20:1500–1506.

Smith HL, Summers BJ, Dillon KH, Macatee RJ, Cougle JR (2016) Hostile interpretation bias in depression. J Affect Disord 203:9–13.

Spunt RP, Lieberman MD (2012) An integrative model of the neural systems supporting the comprehension of observed emotional behavior. NeuroImage 59:3050–3059.

Strangman G, Culver JP, Thompson JH, Boas DA (2002) A quantitative comparison of simultaneous BOLD fMRI and NIRS recordings during functional brain activation. NeuroImage 17:719–731.

Tong Y, Frederick BD (2010) Time lag dependent multimodal processing of concurrent fMRI and near-infrared spectroscopy (NIRS) data suggests a global circulatory origin for low-frequency oscillation signals in human brain. NeuroImage 53:553–564.

Tsuzuki D, Cai D, Dan H, Kyutoku Y, Fujita A, Watanabe E, Dan I (2012) Stable and convenient spatial registration of stand-alone NIRS data through anchor-based probabilistic registration. Neurosci Res 72:163–171.

van Kesteren MTR, Fernández G, Norris DG, Hermans EJ (2010) Persistent schema-dependent hippocampal-neocortical connectivity during memory encoding and postencoding rest in humans. Proc Natl Acad Sci U S A 107:7550–7555.

Van Overwalle F (2009) Social cognition and the brain: a meta-analysis. Hum Brain Mapp 30:829–858.

Verona E, Sadeh N, Curtin JJ (2009) Stress-induced asymmetric frontal brain activity and aggression risk. J Abnorm Psychol 118:131–145.

Wagner DD, Chavez RS, Broom TW (2019) Decoding the neural representation of self and person knowledge with multivariate pattern analysis and data-driven approaches. Wiley Interdiscip Rev Cogn Sci 10:e1482.

Wang Y, Zhu W, Xiao M, Zhang Q, Zhao Y, Zhang H, Chen X, Zheng Y, Xia L-X (2018) Hostile attribution bias mediates the relationship between structural variations in the left middle frontal gyrus and trait angry rumination. Front Psychol 9:526.

Xia M, Wang J, He Y (2013) Brainnet viewer: a network visualization tool for human brain connectomics. PLoS One 8:e68910.

Yang Y, Raine A (2009) Prefrontal structural and functional brain imaging findings in antisocial, violent, and psychopathic individuals: a meta-analysis. Psychiatry Res Neuroimaging 174:81–88.

Yeshurun Y, Swanson S, Simony E, Chen J, Lazaridi C, Honey CJ, Hasson U (2017) Same story, different story: the neural representation of interpretive frameworks. Psychol Sci 28:307–319.

Young L, Bechara A, Tranel D, Damasio H, Hauser M, Damasio A (2010) Damage to ventromedial prefrontal cortex impairs judgment of harmful intent. Neuron 65:845–851.

Yücel MA, Selb J, Cooper RJ, Boas DA (2014) Targeted principle component analysis: a new motion artifact correction approach for near-infrared spectroscopy. J Innov Opt Health Sci 7:1350066.

Yücel MA, Selb JJ, Huppert TJ, Franceschini MA, Boas DA (2017) Functional near infrared spectroscopy: enabling routine functional brain imaging. Curr Opin Biomed Eng 4:78–86.