SUPPLEMENTARY INFORMATION

Manuscript title:

Medial prefrontal and occipito-temporal activity at encoding determine enhanced recognition of threatening faces after 1.5 years

Journal name:

Brain Structure and Function

Authors: Xiqin Liu [1], Xinqi Zhou [1], Yixu Zeng [1], Jialin Li [1], Weihua Zhao [1], Lei Xu [1], Xiaoxiao Zheng [1], Meina Fu [1], Shuxia Yao [1], Carlo V. Cannistraci [2], Keith M. Kendrick [1], Benjamin Becker [1 *]

Affiliations:

[1] Clinical Hospital of Chengdu Brain Science Institute, MOE Key Laboratory for Neuroinformation, University of Electronic Science and Technology of China, Chengdu, China

[2] Center for Complex Network Intelligence (CCNI) at the Tsinghua Laboratory of Brain and Intelligence (THBI), Department of Computer Science, Department of Biomedical Engineering, Tsinghua University, Beijing, China

*** Corresponding author**

Benjamin Becker

Clinical Hospital of Chengdu Brain Science Institute, MOE Key Laboratory for Neuroinformation

University of Electronic Science and Technology of China

Chengdu 611731, China

E-mail: ben_becker@gmx.de, Tel.: Tel.: +86 2861 830 811

SUPPLEMENTARY METHODS

Participants

We re-contacted N = 225 healthy subjects that had participated in a large-scale fMRI project during which they underwent structural, resting-state (Liu et al., 2020) and task-based fMRI assessments (Li et al., 2018; Xu et al., 2020; Zhou et al., 2020). The task-fMRI test battery included blocked design tasks assessing pain empathy (Li et al., 2018; Xu et al., 2020; Zhou et al., 2020) and affective modulation of response inhibition as well as an event-related emotional face processing paradigm. The initial MRI data was acquired between August, 2016 and October, 2017 (Time1, T1) and subjects were re-contacted between July, 2019 and August, 2019 (Time 2, T2) to participate in a surprise recognition memory test for the faces from the emotional face fMRI paradigm.

A total of 102 subjects (53 males, aged 20 – 32 years) agreed to participate in the surprise recognition memory test at T2. All participants were right-handed with normal or corrected-to-normal vision. Due to incomplete behavioral and fMRI data, data from N = 7 participants were discarded; data from N = 4 participants were excluded from further analyses due to extremely low hits and false alarms (hits < 1 & false alarms < 1); N = 2 participants were excluded due to excessive head motion during fMRI scanning (> 3 mm or 3 degrees). Consequently, data from N = 89 subjects was included in the final analyses (44 males, mean age = 23.80 ± 2.39 years, age range: 20-32).

Behavioral data analysis

Univariate approach

A-prime (*A'*) and d-prime (*d'*) both represent bias-free measures of sensitivity or discriminability derived from signal detection theory. Higher *A'* or *d'* scores indicate better discrimination for identifying target faces from lure faces. *A'* was calculated using the following formula: $1/2 + [p(\text{hit}) - p(\text{false alarm})] \times [1 + p(\text{hit}) - p(\text{false alarm})]/\{4 \times p(\text{hit}) \times [1 - p(\text{false alarm})]\}$ (Wang et al., 2012). *d'* is *z-score* difference between the signal and noise distribution $[d' = z(\text{hit}) - z(\text{false alarm})]$. Specifically, in the delayed memory test, *p*(hit) denoted the correct recognition proportion of target faces for which subjects responded "6 (definitely old)", "5 (probably old)" or "4 (maybe old)", whereas *p*(false alarm) denoted the proportion of foil faces for which subjects incorrectly responded "6 (definitely old)", "5 (probably old)" or "4 (maybe old)". In the immediate memory test, *p*(hit) denoted the correct recognition proportion of target faces for which subjects reported "old" and *p* (false alarm) denoted the proportion of foil faces for which subjects incorrectly responded "old". *A'/d'* significantly above chance (0.5/0) confirms that participants can successfully discriminate the target faces from lure faces.

Multivariate approach

PCA was performed to capture the confidence rating pattern of facial expression conditions. The PCA permits dimensionality reduction by extracting groups of covariant features across data and compressing them through an orthogonal linear combination in new variables, which is referred to as principal component (PC). The

PCA analysis is able to provide a new representation of the dataset according to n (where n is equal to the number of samples when it is smaller than the number of features; and it is equal to the number of features when it is smaller than the number of samples) metafeatures called PC scores. Each PCi (with i = 1 to n) score metafeature is a linear combination of the original features according to which we can order the samples, and each PCi explains a percentage of the variance in the original data that is uncorrelated to all other PCi. In particular, the PC1 explains a larger percentage of variance than that of PC2, the PC2 larger than PC3, and so forth till the last component. Compressed data can be mapped in a reduced two-dimensional (2D) space composed by the first two components in order to visualize and discriminate groups of data that are separated. In our study, we assessed whether the PC1 score represents a synthetic meta-participant which provides the highest variance associated to the discrimination of the face images by confidence rating and unsupervisedly mapped the face trials in the 2D geometrical space to visualize whether the trials associated to different facial expressions (angry, fearful, sad, happy and neutral) occupy different zones of the space. The 89 participants' raw confidence rating scores (i.e., features) of all 50 trials (i.e., samples) were considered together to define a multidimensional dataset. Specifically, we conducted the PCA analysis according to the following procedure:

1. The 50 (sample) × 89 (feature) matrix was mapped by PCA into a new reduced 2D geometrical space that was composed by PC1 and PC2. This new 2D geometrical space represented the segregation of the trials according to the highest amount of cumulative explained variance of the dataset.

2. In order to check whether trials associated to different facial expressions occupied different areas of the geometrical space, we attributed a different color to each of the five expression conditions. In practice, we expect that clusters of trials with the same color (facial expression) should dominate different geometrical regions of the 2D space.
3. Next, a non-parametric trustworthiness test was conducted to statistically measure whether the pattern emerging from the amount of variance explained by PC1 is generated at random (Durán et al., 2021). The procedure to compute the trustworthiness exploits a resampling technique based on label-reshuffling. To build a null model distribution, the labels are reshuffled uniformly at random on the embedded points of the PC1 whose location is maintained unaltered in the reduced space. For each random reshuffling, a separability measure value (AUC-ROC in our case) is computed. The collection of all these values is used to draw the null model distribution. This distribution is employed to compute the probability (an empirical *p*-value) to get at random a separation equal or higher than the one detected by using the original labels. If this *p*-value is lower than a significant threshold (that for convention is set to 0.05), then we can trust that the pattern emerging from the amount of variance explained by PC1 is unlikely generated at random.
4. Finally, a pairwise two-tailed Wilcoxon Signed-Rank Test with Benjamini-Hochberg-adjusted correction (for multiple hypothesis testing) was performed for PC1, to statistically assess the extent to which the visible segregation

between groups of trials corresponds to a significant discrimination between different pairs of conditions (e.g., facial expression categories).

fMRI data analysis

Whole-brain two-sample t-test based on subgroups separated by significance of ISC scores

As the subjects can be separated into discriminators and non-discriminators with "qualitatively" different emotional face memory representations based on the significance of the ISC scores, we compared their first-level contrast images using a whole-brain two-sample, two-tailed t-test controlling for interval days using SnPM13. Parameter estimates were extracted from independent masks taken from Brainnetome atlas (http://atlas.brainnetome.org/download.html, Fan et al., 2016) corresponding to the location of the identified clusters using bootstrap tests to warrant a high robustness. Significant clusters in the whole brain were determined using a height threshold of $p < 0.001$ (two-tailed) and an extent threshold of $p < 0.05$ (two-tailed) with cluster-based familywise error (FWE) correction (Eklund et al., 2016; Slotnick et al., 2017).

SUPPLEMENTARY RESULTS

Post-hoc tests on the hit rates in the delayed test

Post-hoc tests with Holm correction for the delayed test revealed that the hit rate of angry faces was significantly higher than that of happy (t_{88} = 5.05, p < 0.001), neutral (t_{88} = 4.84, p < 0.001) and sad (t_{88} = 3.31, p < 0.01) faces, whereas there was no difference between angry and fearful faces (t_{88} = 1.09, p = 0.558). The hit rate of fearful faces was also significantly higher than that of happy (t_{88} = 5.43, p < 0.001), neutral (t_{88} = 4.97, p < 0.001) and sad (t_{88} = 2.79, p < 0.05) faces. Happy and neutral faces was not different in hit rates (t_{88} = 0.10, p = 0.923), while sad faces were significantly better remembered than neutral faces (t_{88} = -2.63, p = 0.04) and marginally significantly better remembered than happy faces (t_{88} = -2.37, p = 0.06) and after Holm correction.

Correlation analyses between hit rates and interval days in the delayed test

The correlations between interval days and hit rate in each condition were not significant (angry: r = 0.17, p > 0.05; fearful: r = -0.03, p > 0.05; happy: r = 0.13, p > 0.05; neutral: r = -0.14, p > 0.05; sad: r = -0.02, p > 0.05).

One-sample t test examining chance level of hit rates and false alarm rates at immediate and delayed test

In the delayed test, angry (t_{88} = 3.00, p < 0.005) and fearful (t_{88} = 2.16, p < 0.05) faces were recognized higher than chance level while the hit rates for neutral (t_{88} =-2.89, p = 0.005) and happy (t_{88} =-3.09, p < 0.005) faces were lower than chance level, and the hit rate for sad faces was at chance level (t_{88} = -0.45, p = 0.66). In the

immediate test, the hit rates for all expression conditions were higher than chance level (angry: t_{88} =5.56, p < 0.001; fearful: t_{88} =5.07, p < 0.001; happy: t_{88} =3.09, p < 0.005; neutral: t_{88} =4.97, p < 0.001; sad: t_{88} =3.49, p < 0.005). For false alarm rates, while the angry (t_{88} = 0.70, p = 0.49) and fearful (t_{88} = 1.61, p = 0.11) faces in the delayed test were falsely recognized by chance, the false alarm rates for neutral (t_{88} =-3.59, p < 0.005), happy (t_{88} =-4.97, p < 0.001) and sad (t_{88} =-2.43, p < 0.05) were lower than chance level. In the immediate test, the false alarm rates for angry (t_{88} =-0.32, p = 0.75) and neutral (t_{88} =-0.19, p = 0.85) faces were at chance level, whereas those for fearful (t_{88} =-5.47, p < 0.001), happy (t_{88} =-5.68, p < 0.001) and sad (t_{88} =-7.68, p < 0.001) faces were less than chance level.

Post-hoc tests on the false alarm rates in the immediate test

Post-hoc tests with Holm correction for the immediate test revealed that the false alarm rate of angry faces was significantly higher than that of fearful (t_{88} = 5.63, p < 0.001), happy (t_{88} = 5.15, p < 0.001) and sad (t_{88} = 7.18, p < 0.001) faces. The false alarm rate of neutral faces was also significantly higher than that of fearful (t_{88} = 4.11, p < 0.001), happy (t_{88} = 5.47, p < 0.001) and sad (t_{88} = 6.33, p < 0.001) faces. By contrast, there was no difference between angry and neutral faces (t_{88} = -0.11, p = 1.00, fearful and happy faces (t_{88} = 0.55, p = 1.00), fearful and sad faces (t_{88} = 2.13, p = 0.14), happy and sad faces (t_{88} = 1.11, p = 0.81) after Holm correction.

Post-hoc tests on the subjective arousal ratings in the current sample

Post-hoc tests with Holm correction for the subjective arousal ratings in the current sample revealed that the arousal rating of neutral faces was significantly lower

than that of angry faces (t_{88} = -11.53, p < 0.001), fearful faces (t_{88} = -10.98, p < 0.001), happy faces (t_{88} = -12.20, p < 0.001) and sad faces (t_{88} = -9.46, p < 0.001). The arousal ratings of angry (t_{88} = 5.25, p < 0.001), fearful (t_{88} = 4.28, p < 0.001) and happy (t_{88} = 2.86, p < 0.05) faces were also significantly higher than that of sad faces. By contrast, there was no difference between angry and fearful faces (t_{88} = 1.46, p = 0.44), angry and happy faces (t_{88} = 1.40, p = 0.44), fearful and happy faces (t_{88} = 0.38, p = 0.70).

Post-hoc of Wilcoxon Signed-Rank tests on the confidence ratings in the delayed test

Non-parametric Wilcoxon Signed-Rank test showed that the confidence ratings for angry and fearful faces were not different (Wilcoxon sign-rank z = -1.33, p = 0.18). The ratings for happy and neutral faces also did not differ (Wilcoxon sign-rank z = -0.11, p = 0.91). On the other hand, there were significant differences of the confidence rating between facial expressions angry vs. happy (Wilcoxon sign-rank z = -5.25, p < 0.001), angry vs. neutral (Wilcoxon sign-rank z = -5.05, p < 0.001) angry vs. sad (Wilcoxon sign-rank z = -3.82, p < 0.001), fear vs. happy (Wilcoxon sign-rank z = -5.20, p < 0.001), fear vs. neutral (Wilcoxon sign-rank z = -5.32, p < 0.001), fear vs. sad (Wilcoxon sign-rank z = -2.75, p = 0.006), happy vs. sad (Wilcoxon sign-rank z = -2.11, p = 0.035) and neutral vs. sad (Wilcoxon sign-rank z = -2.55, p = 0.011). The results remained stable after Benjamini-Hochberg adjustment for multiple comparison testing.

Post-hoc of Wilcoxon Signed-Rank tests on the old/new forced choice in the immediate test

Non-parametric Wilcoxon Signed-Rank test showed that the old/new judgment for angry and fearful faces were not different (Wilcoxon sign-rank z = -1.33, p = 0.18). The ratings for happy and neutral faces also did not differ (Wilcoxon sign-rank z = -0.11, p = 0.91). On the other hand, there were significant differences of the confidence rating between facial expressions angry vs. happy (Wilcoxon sign-rank z = -5.25, p < 0.001), angry vs. neutral (Wilcoxon sign-rank z = -5.05, p < 0.001) angry vs. sad (Wilcoxon sign-rank z = -3.82, p < 0.001), fear vs. happy (Wilcoxon sign-rank z = -5.20, p < 0.001), fear vs. neutral (Wilcoxon sign-rank z = -5.32, p < 0.001), fear vs. sad (Wilcoxon sign-rank z = -2.75, p = 0.006), happy vs. sad (Wilcoxon sign-rank z = -2.11, p = 0.035) and neutral vs. sad (Wilcoxon sign-rank z = -2.55, p = 0.011). The results remained stable after Benjamini-Hochberg adjustment for multiple comparison testing.

Distinct long-term face memory patterns between subgroups

The BPSA approach successfully separated the subjects into discriminators (N = 43, 21 males, rs > 0.28, ps < 0.05) and non-discriminators (N = 46, 23 males, rs < 0.28, ps > 0.05). To further confirm the separable emotional face representations between the two subgroups, we calculated the similarity between each discriminator's multi-item discriminability pattern and the mean pattern of the discriminators (within-group similarity) and compared it with the similarity between each non-discriminator's multi-item discriminability pattern and the mean pattern of the discriminators (between-group similarity). As expected, the within-discriminator

group similarity was significantly higher than between-group similarity (t_{87} = 14.20, p < 0.001, Cohen's d = 3.01), suggesting that the non-discriminators' discriminative patterns significantly deviated from that of the discriminator group. The distinct memory pattern between the two groups was further confirmed using a univariate method by comparing the confidence ratings of the angry/fearful (bin), sad and happy/neutral (bin) faces, given that both univariate and multivariate PCA results suggested a separation of recognition performance among the three conditions. A significant 2-way interaction was detected ($F_{(2,86)}$ = 20.49, p < 0.001, η^2 = 0.19, **Fig. 4A**).

Post-hoc tests showed significant differences between all pairs of the expression conditions in the discriminators, but no differences in the non-discriminators. Specifically, the averaged confidence ratings for faces with angry/fearful expressions were higher compared to sad (t_{42} = 5.54, corrected p < 0.001, Cohen's d = 0.84) and happy/neutral expressions (t_{42} = 10.32, corrected p < 0.001, Cohen's d = 1.57), and ratings for sad were higher than happy/neutral expressions (t_{42} = 3.98, corrected p < 0.001, Cohen's d = 0.61) in the discriminative group, whereas there was no difference in the non-discriminative group (angry/fearful vs. sad: t_{45} = 0.79, p = 0.43; angry/fearful vs. happy/neutral: t_{45} = 1.48, p = 0.15; sad vs. happy/neutral: t_{44} = 0.35, p = 0.73). Further control analyses suggested that the discriminative and non-discriminative group did not differ with respect to age (t_{87} = 0.33, p = 0.74), gender distribution ($\chi^2_{(1)}$ = 0.01, p = 0.92), interval days (t_{87} = -1.53, p = 0.13), arousal ratings for each expression condition ($F_{(4,84)}$ = 0.59, p = 0.67) and the overall recognition

performance for all faces (t_{87} = 1.27, p = 0.21), arguing against confounding effects of these variables. Moreover, the different memory patterns among angry/fearful, sad and happy/neutral of the two groups in the delayed test was not present at immediate recognition (interaction on hit rates: $F_{(2,86)}$ = 0.24, p = 0.79). The distinguishable response patterns of the two subgroups thus allowed us to conduct the fMRI analysis with sufficient power by group comparison.

Whole-brain two-sample t-test on the group separation based on ISC scores

To further validate the findings, two-sample t-test between the discriminators and non-discriminators with distinct memory pattern on emotional expression conditions were conducted which revealed significant interaction effects in the left IOG (k = 127, peak MNI coordinates: -18, -82, -19; t = -4.65; two-tailed $p_{\text{cluster-FWE}} < 0.05$, whole-brain corrected) and right vmPFC/OFC (k = 131, peak MNI coordinates: 27, 32, -16; t = -3.90; two-tailed $p_{\text{cluster-FWE}} < 0.05$, whole-brain corrected; **Fig. 5A**), consistent with the brain-behavior correlation results with the overlap coefficients (defined as the size of intersection divided by the smaller of the two clusters, Saygin et al., 2016) of 0.60 and 0.53 in the left IOG, right vmPFC/OFC, respectively. To further disentangle the effects in group comparison, post-hoc analyses were performed with independent masks that correspond to the location of the left IOG (atlas label 205, cluster size = 271, overlapping voxels = 42; **Fig. 5A**) and right vmPFC/OFC (atlas label 46, cluster size = 356, overlapping voxels = 64; **Fig. 5A**) from the Brainnetome atlas (Fan et al., 2016). Extraction of beta estimates from the masks showed that the discriminators responded

with lower activation for threatening vs. non-threatening faces than non-discriminators in both the left IOG (post-hoc $p < 0.001$, two-tailed Bootstrap test, **Fig. S3**) and the right vmPFC/OFC (post-hoc $p < 0.001$, two-tailed Bootstrap test, **Fig. S3**). A further extraction of the estimates of the threatening and non-threatening face conditions respectively from the masks suggested a distinct pattern of neural responses in the left IOG and right vmPFC/OFC. That is, in the left IOG, the non-discriminators showed higher reactivity towards threatening faces compared to non-threatening faces (t_{45} = 4.50, $p < 0.001$, Holm corrected, Cohen's d = 0.66) while there was no difference in discriminators (t_{42} = -0.49, $p = 0.63$, **Fig. 5B**). Conversely, the discriminators responded less to threatening faces compared to non-threatening faces (t_{42} = -3.66, $p < 0.01$, Holm corrected, Cohen's d = 0.56) whereas the non-discriminators showed no difference in the right vmPFC/OFC (t_{45} = 1.72, $p = 0.09$, **Fig. 5C**). Additionally, examining the main effects of the emotional expression and group revealed that the bilateral middle temporal gyrus (right MTG: $k = 295$, peak MNI coordinates: 60, -55, -1, $t = 5.52$, $p_{\text{cluster-FWE}} < 0.05$, whole-brain corrected; left MTG: $k = 189$, peak MNI coordinates: -54, -58, 2, $t = 4.50$, $p_{\text{cluster-FWE}} < 0.05$, whole-brain corrected, **Fig. S4**) and the left fusiform gyrus ($k = 130$, peak MNI coordinates: -33 -73, -16, $t = 4.29$, $p_{\text{cluster-FWE}} < 0.05$, whole-brain corrected, **Fig. S4**) exhibited higher reactivity towards threatening vs. non-threatening faces during encoding irrespective of group, whereas no significant main effect of group was observed, arguing against unspecific face encoding differences between the groups.

SUPPLEMENTARY FIGURES

Fig. S1 Subjective arousal ratings for each emotional expression condition in the current sample (n = 89). Error bars depict ±1 SEM

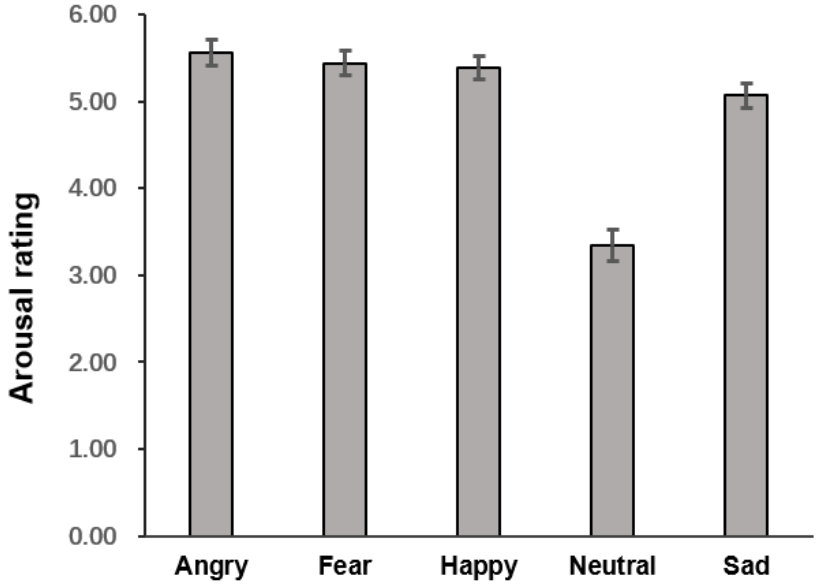


Fig. S2 PCA analysis of forced old/new responses in the immediate test. Facial expression conditions were not separated along PC1 axis according to the scatter plot and the distributions. Wilcoxon Signed-Rank tests also detected no significant differences between facial expression conditions (all *ps* > 0.05)

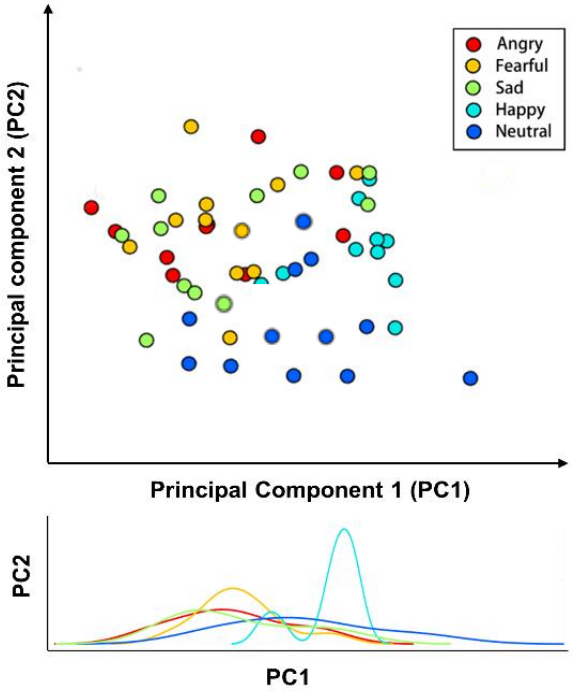


Fig. S3 Post-hoc tests on the extracted parameter estimates. Extraction of parameter estimates from the left IOG and right vmPFC/OFC (beta estimates) as defined by independent masks from the Brainnetome atlas showed that discriminators responded with lower activation for threatening vs. non-threatening faces than non-discriminators in both the left IOG and the right vmPFC/OFC

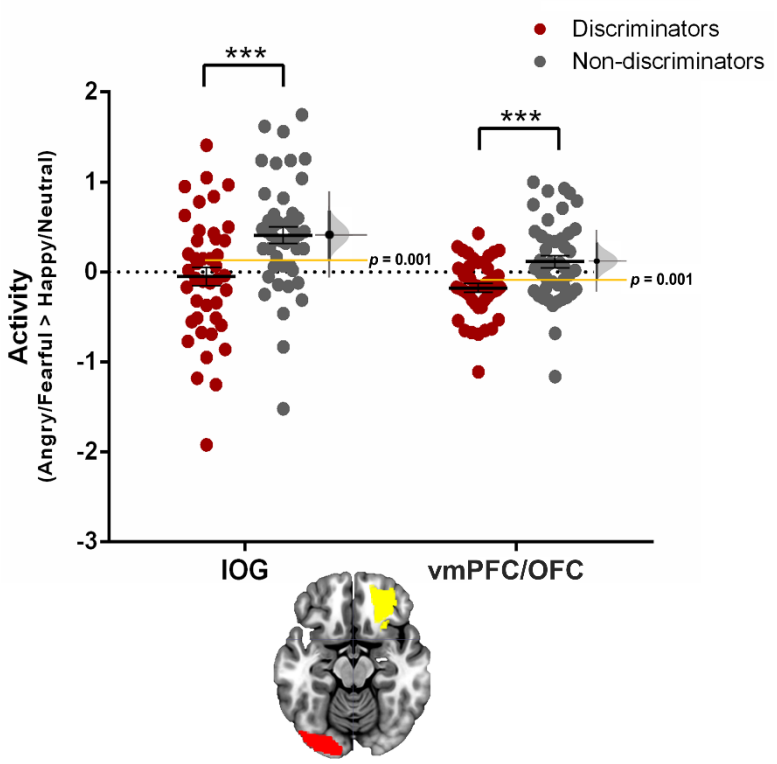
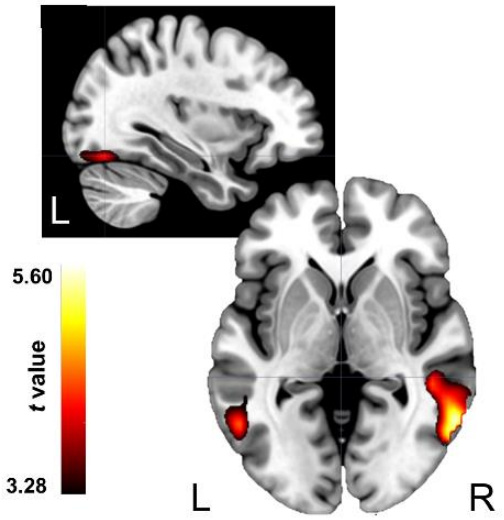


Fig. S4 Main effect of the emotional expression revealed the bilateral middle temporal gyrus (MTG) and left fusiform gyrus showing higher neural responses to threatening vs. non-threatening faces during encoding. Statistical image is displayed at $p < 0.05$ cluster-level FWE correction with a cluster-forming threshold $p < 0.001$



Supplementary references

Durán, C., Ciucci, S., Palladini, A., Ijaz, U. Z., Zippo, A. G., Sterbini, F. P., … Cannistraci, C. V. (2021). Nonlinear machine learning pattern recognition and bacteria-metabolite multilayer network analysis of perturbed gastric microbiome. *Nature Communications, 12*(1). https://doi.org/10.1038/s41467-021-22135-x

Li, J., Xu, L., Zheng, X., et al. (2019). Common and Dissociable Contributions of Alexithymia and Autism to Domain-Specific Interoceptive Dysregulations: A Dimensional Neuroimaging Approach. *Psychotherapy and Psychosomatics, 88*(3), 187-189.

Liu, C., Xu, L., Li, J., Zhou, F., Yang, X., Zheng, X., … Becker, B. (2020). Serotonin and early life stress interact to shape brain architecture and anxious avoidant behavior-A TPH2 imaging genetics approach. *Psychological Medicine*. https://doi.org/10.1017/S0033291720002809

Xu, L., Bolt, T., Nomi, J. S., Li, J., Zheng, X., Fu, M., … Uddin, L. Q. (2020). Inter-subject phase synchronization differentiates neural networks underlying physical pain empathy. *Social Cognitive and Affective Neuroscience, 15*(2), 225–233. https://doi.org/10.1093/scan/nsaa025

Zhou, F., Li, J., Zhao, W., Xu, L., Zheng, X., Fu, M., … Becker, B. (2020). Empathic pain evoked by sensory and emotional-communicative cues share common and process-specific neural representations. *ELife, 9*, 1–27. https://doi.org/10.7554/ELIFE.56929