Title: Medial prefrontal activity at encoding determines recognition of threatening faces after 1.5 years

Running title: Long-lasting memory for facial threats

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, 3], 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] Biomedical Cybernetics Group, Biotechnology Center (BIOTEC), Center for Molecular and Cellular Bioengineering (CMCB), Center for Systems Biology Dresden (CSBD), Cluster of Excellence Physics of Life (PoL), Department of Physics, Technische Universität Dresden, Dresden, Germany

[3] Center for Complex Network Intelligence(CCNI), Tsinghua Laboratory of Brain and Intelligence (THBI), Department of Bioengineering, 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

Abstract

Studies demonstrated that faces with emotional expressions are better remembered than neutral faces. However, whether the memory advantage persists over years and which neural systems mediate such an effect remains unknown. We investigated recognition of incidentally encoded faces with angry, fearful, happy, sad and neutral expressions over >1.5 years (N=102). Both univariate and multivariate analyses showed that faces with threatening expressions (angry, fearful) were better recognized than happy and neutral faces. Comparison with immediate recognition indicated that this effect was driven by decreased recognition of non-threatening faces. Functional magnetic resonance imaging (fMRI) data was acquired during encoding and results revealed that differential neural encoding in the right ventromedial prefrontal/orbitofrontal cortex neurally mediated the long-term recognition advantage for threatening faces. Our study provides the evidence that threatening facial expressions lead to persistent face recognition over periods of >1.5 years and encoding-related activity in the ventromedial prefrontal cortex may underlie preserved recognition.

Keyword: Face recognition; emotion; threat; fMRI; encoding; medial prefrontal cortex

1. Introduction

For social species the recognition of previously encountered conspecifics is vital for survival and successful interaction. In humans, faces are presumably the most important stimuli for subsequent recognition. Given the high evolutionary significance of these stimuli, cortical networks specialized in perceiving and recognizing faces develop already during early infancy (Powell et al., 2018, Cohen et al., 2019). Nevertheless, the ability to recognize faces varies greatly in the human population. While some individuals can recognize faces following a single exposure over years, others find it nearly impossible to recognize highly familiar faces (Russell et al., 2009; Tardif et al., 2019). In addition to individual differences, several characteristics of the facial stimuli can affect subsequent recognition including emotional expression (Bruce and Young, 1986; Haxby et al., 2000). From an evolutionary perspective, the emotional expression may transmit important information such that threat-associated facial expressions during initial encounters may relate to harm avoidance in future encounters (Darvin, 1872; Staugaard, 2010).

In support of this evolutionary hypothesis, some experimental studies have demonstrated that faces with threatening expressions (e.g., angry or fearful) are both better detected and remembered as compared to those with non-threatening expressions (i.e. neutral, sad or happy, Grady et al., 2007; Jackson et al., 2014; Keightley et al., 2011; Stiernströmer et al., 2016; Wang, 2013). These findings broadly align with numerous previous studies that demonstrated a robust emotional enhancement of memory for non-facial stimuli indicating that emotional events, particularly high-arousing negative ones,

are more vividly and accurately remembered over retention intervals ranging from minutes to years (reviewed in Bowen et al., 2018; Yonelinas and Ritchey, 2015). In contrast, examination of specific effects of emotional facial expressions on subsequent recognition revealed inconclusive results and findings varied according to retention intervals (e.g., Anderson et al., 2006; Gupta and Srinivasan, 2009; Mather and Carstensen, 2003; Pinabiaux et al., 2013; Wang, 2013).

While studies examining effects in the domain of working memory consistently reported better recognition of face with threatening expressions (Jackson et al., 2014; Öhman et al., 2001; Thomas et al., 2014), findings regarding a short-term or long-term memory advantage of threatening faces were less clear. For instance, a number of studies showed no effect of face expression on memory performance immediately after encoding (Grady et al., 2007; Satterthwaite et al., 2009; Xiu et al., 2015) whereas others reported better recognition for fearful faces compared to neutral faces after a 24-h delay (Wang, 2013). On the other hand, Anderson et al. (2006) found that fearful faces were not subject to enhanced recognition at retention intervals ranging from 15 min to 2 weeks. Most importantly, effects of emotional facial expressions on recognition after longer retention intervals (i.e., years) and the underlying neural basis have not been systematically examined.

Against this background, the present study investigated the long-term emotional expression effects by capitalizing on a large fMRI study during which $N = 225$ healthy young adults underwent incidental encoding of faces with emotional expressions (angry, fearful, happy, sad and neutral), of whom $N = 102$ participated in a surprise face

recognition test scheduled at least 1.5 years after encoding. We tested whether face recognition was modulated by facial expressions by means of univariate and innovative data-driven multivariate approaches which allowed us to capitalize on the entire response pattern of the subjects. Based on the extensive previous literature reporting robustly enhanced recognition of non-facial negative stimuli across long-term memory retention intervals ranging from days (Anderson et al., 2006; Ritchey et al., 2011;Wang, 2018; Wang et al., 2014) to 1 year (Dolcos et al., 2005; Erk et al., 2010; Gavazzeni et al., 2012), we hypothesized augmented recognition of faces with threatening expressions, particularly angry and fearful, after a retention interval of >1.5 years (hypothesis 1).

To further explore which neural systems mediated the expression-associated memory advantage, we examined whether brain activation during incidental encoding varied as a function of expression-specific memory performance after 1.5 years. To this end, we employed an innovative multivariate pattern similarity analysis of behavioral response (BPSA) to classify participants who exhibited an expression-specific memory advantage and participants who did not express such an effect. We next explored differential encoding-related activity between the classified participants on the whole brain-level as well as in the amygdala. Both emotional face processing and emotional-enhancement of memory have been strongly linked to limbic and prefrontal systems, specifically regions engaged in emotional reactivity and value processing such as the medial prefrontal cortex (mPFC) and amygdala (Hariri et al., 2003; Kark and Kensinger, 2019; Kensinger and Schacter, 2008; Ritchey et al., 2011; Tyng et al., 2017; Vuilleumier

et al., 2004). Thus, on the neural level we conjectured that encoding-related activity in these systems may mediate individual differences in the long-term effects of emotional expressions on face recognition (hypothesis 2).

2. Methods

2.1 Participants

A total of 102 healthy, young right-handed Chinese students participated in this study which was part of a large-scale fMRI project (e.g., Li et al., 2018; Liu et al., 2019; Xu et al., 2020; Zhou et al., 2020). Following initial quality assessment data from N = 83 subjects was included in the final analyses (40 males; mean age = 21.40 ± 2.40 years, age range: 18-30). Details on recruitment protocols and quality assessments are provide in the **Supplementary Methods**. The study was approved by the local ethics committee at the University of Electronic Science and Technology of China and in accordance with the latest revision of the Declaration of Helsinki. Written informed consent was obtained from each participant.

2.2 Stimuli

A total of 150 face stimuli were selected from two validated Asian facial expression databases: Chinese Facial Affective Picture System (Gong et al., 2011) and Taiwanese Facial Expression Image Database (TFEID) (Chen and Yen, 2007). Facial expressions included angry, fearful, sad, happy and neutral (each from 30 different individual actors, 15 males). All facial stimuli were gray-scaled and covered with an oval mask to remove individual features (e.g., hair). The 150 faces were evenly divided into three sets

matched with regard to arousal and valence for the current experiment. Within each face set, the arousal ratings of emotional faces (angry, fearful, happy, sad) were higher as compared to neutral faces (all *ps* < 0.001), whereas arousal ratings between emotional faces did not differ (all *ps* > 0.05).

2.3 Experimental procedure

The present study employed a multiple-stage procedure including an incidental encoding phase and a subsequent memory phase (Fig.1). All participants initially underwent an event-related fMRI paradigm using an emotional face processing task (i.e., incidental encoding) between August, 2016 and October, 2017 (Time 1, T1). Two runs of the facial stimuli (50 stimuli per run, set 1), balanced for facial expression and gender, were presented (5min 12s per run). Stimuli were shown for 2500ms during which the participants were required to judge the gender of the face by button press. After each trial, a jittered fixation cross was presented for 2000–5600ms (mean ITI = 3800ms, **Fig. 1**). Stimuli were presented via E-prime 2.0 (Psychology Software Tools, USA, http://www.pstnet.com/eprime.cfm). Twenty minutes after fMRI acquisition, participants were asked to complete a surprise recognition memory test (immediate test) outside the scanner in which the 50 previously presented faces (set 1, targets) from the fMRI paradigm were intermixed with 50 new faces (set 2, lures). Participants were instructed to indicate whether each face had been shown during the fMRI acquisition (forced choice: old versus new). After a retention interval of >1.5 year (interval range: 653-1113 days), participants were requested to perform a surprise recognition test (delayed test) between July, 2019 and August, 2019 (Time 2, T2) in which target faces

were intermixed with another set of 50 new faces (set 3, lures). In the delayed test, participants were asked to rate their recognition confidence on a six-point scale (old vs. new; 1 = definitely new to 6 = definitely old, Fig. 1). The confidence rating approach was employed given that it reflects the strength and quality of the memory more precisely (Aly and Turk-Browne, 2016; Stretch and Wixted, 1998), and thus was more sensitive as compared to the categorical old/new judgement approach in a long retention interval. Moreover, this allowed us to conduct multivariate analysis on the delayed test data thus increasing power. The delayed recognition memory test was carried out online via SurveyCoder 3.0 (https://www.surveycoder.com/).

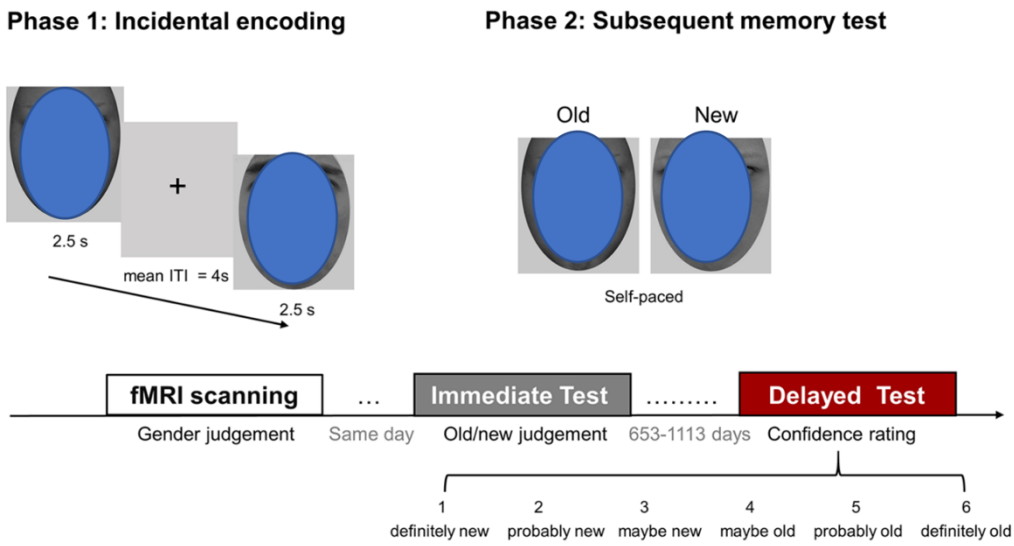

Fig. 1. Experimental design and stimuli. Upper panel: Face stimuli in the encoding and subsequent memory stage (note faces have been covered with blue ovals for the schematic presentation in the paper). Lower panel: Experimental procedure and the behavioral measures.

2.4 MRI data Acquisition

MRI data were obtained on a 3T GE MRI system (General Electric, Milwaukee, WI, USA) using standard sequence parameters (see **Supplementary Methods**).

2.5 data analysis

2.5.1 Univariate approach

To examine effects of facial expressions on recognition, hit rates were entered in a two-way analysis of variance (ANOVA) with facial expression (angry, fearful, happy, neutral and sad) and retention interval (immediate and delayed test) as within-subject factors. For both retention intervals, hit rates were defined as the ratio of target faces correctly identified as old. For the delayed test, ratings of 4, 5 and 6 were considered as correctly identified. Post-hoc tests for significant interactions with Holm-Bonferroni correction were conducted to examine the expression-associated immediate and delayed memory effects, and planned two-tailed t tests were performed to determine changes of memory in each facial expression condition over >1.5 years. Notably, the analyses focused primarily on hit rates because: (*i*) it allows a direct comparison of immediate and delayed test performance, and *(ii)* it focused on the corrected responses of target faces thus establishing a link between the recognition performance and the encoding process, aligning with the goal of the present study which was to reveal the neural basis during encoding that might contribute to long-term emotional face recognition. Analyses were computed in SPSS (IBM, SPSS version 20, 2011).

2.5.2 Multivariate approach

Given that the hypothesis-driven univariate analysis (e.g., ANOVA) might dismiss

individual differences related to response variability due to averaging, we further investigated the robustness of long-term emotional memory advantage using a data-driven unsupervised (the information emerges from the data without any constraint on our expectation) multivariate approach. The schematic of multivariate analyses is shown in **Fig. 2**. To this end, principal components analysis (PCA) was implemented to capitalize on the full pattern of behavioral responses (**Fig. 2A, 2B**). PCA is one of the most widely used unsupervised multivariate machine learning algorithms for exploring the hidden pattern of multidimensional data (Ringnér, 2008), which allowed us to visualize similarities and differences in a data set. Consistent with the univariate analysis, the multivariate PCA only focused on the confidence ratings of targets in this study. Specifically, the confidence ratings of all target face trials were plotted in a two-dimensional reduced space composed by the first (PC1) and second principal component (PC2) (details see **Supplementary Methods**). PC1 explains the highest variance and represents the most discriminative dimension. We thus expected that trials with the same facial expression (color coded) would dominate separate regions within the reduced space along the PC1 axis, representing a discriminative emotion-specific face memory pattern (**Fig. 2B**). Pairwise two-tailed Wilcoxon Signed-Rank Tests with Benjamini-Hochberg-adjusted correction were further performed to statistically assess the separation. Multivariate analyses were implemented in PC-corr MATLAB code (https://github.com/biomedical-cybernetics/PC-corr_net) which has been used in previous studies to successfully discriminate behavioral and omic patterns (Miendlarzewska et al., 2018; Ciucci et al., 2017). The multivariate PCA combined with

the univariate approach thus facilitated a sensitive determination of emotional expression-specific face recognition memory over an extensive retention interval.

2.5.3 Participant grouping – Behavioral response pattern similarity analysis (BPSA)

For the current analysis, we integrated the inter-subject correlation analysis (e.g., Cantlon and Li, 2013; Tian et al., 2019) with pattern similarity analysis (e.g., Aly and Turk-Browne, 2016; Günseli and Aly, 2020) towards an innovative approach that is based on the similarity between individual behavioral response patterns and their principal component derived from PCA (termed "behavioral response pattern similarity analysis", BPSA). This approach aims to identify subjects who exhibited an emotion-specific discriminative memory pattern from those who did not.

To do this, we extracted the PC1 loadings as a response template (i.e., PC1 score) and then calculated Pearson's correlation coefficients between each participant's confidence rating pattern for targets and this PC1 response template (**Fig. 2C**). Based on the significance of the correlations, we categorized the participants into two groups. Participants whose correlation coefficients were significant ($p < 0.05$) were classified as discriminative group exhibiting an emotion-specific memory effect, whereas those with non-significant similarities were classified as non-discriminative group without exhibiting such an effect (**Fig. 2D**). The pattern similarity-based group definition was used to inform the subsequent analysis of the fMRI data that aimed at determining the neural basis of long-term emotional face memory with sufficient statistical power.

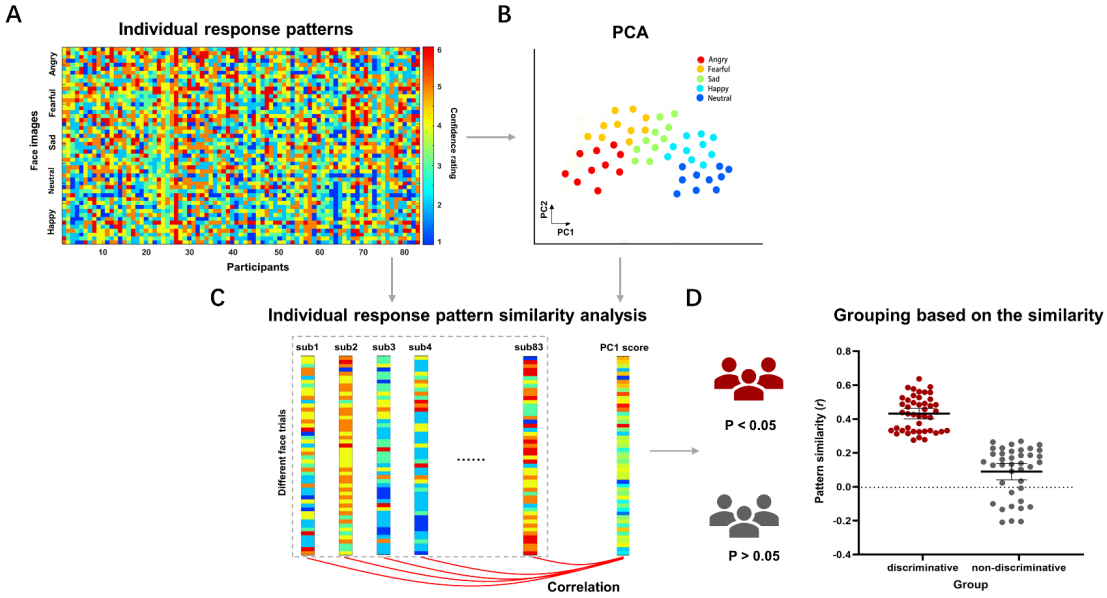

Fig. 2. Schematic of major steps in multivariate investigation of the long-term emotional face recognition memory. (*A*) Representation of confidence rating response patterns for participants. Each column represents the response pattern for each participant. Rows from top to bottom indicate face images with expressions angry, fearful, sad, neutral and happy. Color blue to yellow designates the confidence rating responses from 1 to 6. (*B*) PCA dimensionality reduction and expected discriminative pattern of the facial expression conditions with sample plotted in the 2D reduced space. (*C*) Individual response pattern similarity analysis between each participant's response pattern and the PC1 template derived from PCA. (*D*) Expected grouping results of participants based on the individual response pattern similarity analysis (i.e., one subgroup exhibits discriminative response pattern for facial expression conditions, and one subgroup does not.)

2.6 fMRI data analysis

2.6.1 Image preprocessing

The functional MRI data was preprocessed and analyzed using SPM12 (Statistical Parametric Mapping, https://www.fil.ion.ucl.ac.uk/spm/software/spm12/). Data preprocessing details are presented in **Supplementary Methods**.

2.6.2 Statistical analyses

To identify the neural substrates associated with the long-term emotional expression effects on face memory, a whole-brain ANOVA model with emotional expression (e.g., threatening vs. non-threatening faces determined by the behavioral results) as within-subject factor, and group (discriminative vs. non-discriminative determined by the pattern similarity analysis) as between-subject factor was employed. To this end, the first-level contrast of interest (e.g., threatening vs. non-threatening) was modeled and subjected to a two-sample, two-tailed t-test comparing the discriminative and non-discriminative group. This analysis allowed us to identify the specific regions sensitive to the long-term emotional memory advantage during encoding within the discriminative group while controlling for unspecific processes by comparison with the non-discriminative group. A whole-brain threshold with $p < 0.05$ Family-Wise Error (FWE) correction at the cluster-level and an initial cluster-forming threshold of $p < 0.001$ was employed (Eklund et al., 2016; Slotnick et al., 2017).

Given that the amygdala has been strongly implicated in emotion processing and emotional memory formation (Blanchard and Blanchard ,1972; Davis, 1992; LaBar and Cabeza, 2006; LeDoux,1995), we examined effects in the amygdala with increased sensitivity using *a priori* region of interest (ROI) analysis. ROIs were based on the Automated Anatomic Labelling (AAL) atlas (Tzourio-Mazoyer et al., 2002). Small volume correction was performed using FWE correction with a voxel-level threshold of $p < 0.05$.

3. Results

Overall, participants successfully discriminated target faces from lure faces in both the immediate test (mean *A'* = 0.65, $p < 0.001$) and the delayed test (mean *A'* = 0.54, p < 0.001). Paired t-test further indicated that *A'* in the delayed test was significantly lower than that in the immediate test ($t_{82} = -8.99$, $p < 0.001$, Cohen's d = 1.21), suggesting that the general recognition performance decreased with time.

3.1 Long-term emotional expression effects on face memory

3.1.1 Univariate results

The ANOVA revealed a significant main effect of facial expression ($F_{(4,79)} = 6.77$, $p < 0.001$, $\eta^2 = 0.26$) and retention interval ($F_{(4,82)} = 23.60$, $p < 0.001$, $\eta^2 = 0.22$), as well as a significant interaction effect ($F_{(4,79)} = 4.86$, $p < 0.005$, $\eta^2 = 0.20$, **Fig. 3**). Post-hoc tests indicated that facial expression modulated delayed memory performance ($F_{(4,79)} = 13.30$, $p < 0.001$, $\eta^2 = 0.40$), such that hit rates for faces with both threatening facial expressions (angry or fearful) were significantly higher as compared to faces with non-threatening expressions (neutral or happy, respectively) (two-tailed *ps* < 0.001, Holm-Bonferroni corrected). In contrast, no significant differences between recognition performance of angry versus fearful as well as neutral versus happy faces were observed, whereas that of sad faces ranged in between (see **Supplementary Results**). To rule out the possibility that the long-term emotional memory advantage of threatening faces was influenced by variations in the retention interval (ranging from 653-1113 days), the repeated-measures ANOVA was recomputed including interval day as a covariate and results remained robust ($F_{(4,78)} = 2.95$, $p = 0.025$, $\eta^2 = 0.13$).

During immediate recognition, a marginally significant difference of hit rates for the five facial expressions was observed ($F_{(4,79)} = 2.43$, $p = 0.055$, $\eta^2 = 0.11$), whereas the post-hoc comparisons revealed no difference between any pairs after correction (see **Supplementary Results**). Pairwise comparisons for each facial expression condition between the immediate and delay test suggested that recognition performance for happy, sad and neutral faces (two-tailed paired t-test: happy: $t_{82} = -5.08$, $p < 0.001$, Cohen's d = 0.44; sad: $t_{82} = -2.59$, $p = 0.012$, Cohen's d = 0.28; neutral: $t_{82} = -5.97$, $p < 0.001$, Cohen's d = 0.64, Fig. 3) significantly declined during the 1.5-year retention interval. The results remained significant after Holm-Bonferroni correction for multiple comparison testing.

Together, the results indicated a long-term face recognition advantage of threatening expressions (i.e., angry and fearful) and this advantage was driven by decreased recognition of faces with non-threatening expressions including happy, sad and neutral following a retention interval of 1.5 years.

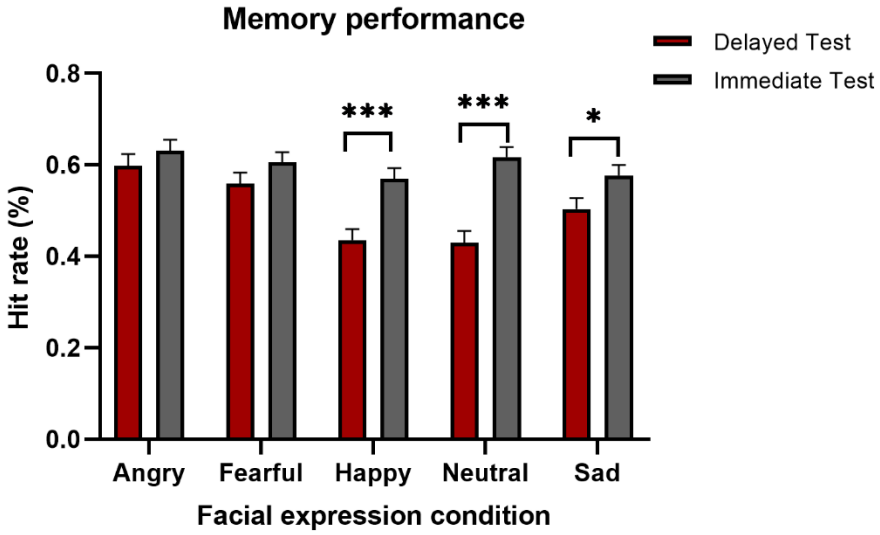

Fig. 3. Memory performance (hit rate) for each facial expression condition in the immediate and delayed recognition memory test. Error bars depict ±1 SEM. *** $p < 0.001$, Holm-Bonferroni corrected.

3.1.2 Multivariate results

Initial inspection of the response patterns (color-coded trial-wise confidence ratings with blue to yellow as the confidence increased) revealed strong individual variations in the confidence ratings for expression-specific face images after long retention intervals (**Fig. 2A**). PCA was then applied to map the confidence ratings of all 50 target face images in the 2D geometrical space of PC1 and PC2. As expected, PC1 has a discriminative variability that accounts for facial expression conditions (**Fig. 4A**). In particular, considering the localization of each face trial along the PC1 axis as visual reference, angry and fearful were clearly separated from neutral and happy while sad was localized in between (**Fig. 4B**). PC1 explained 12.10% of the variance. In line with the visual presentation, pairwise non-parametric Wilcoxon Signed-Rank tests revealed significant differences between the facial expression conditions except for angry vs.

fearful and neutral vs. happy (see **Supplementary Results**). The same PCA procedure was also applied to immediate test and failed to detect a discriminative memory pattern (see **Supplementary Figure S1**).

To summarize, both univariate and multivariate analyses suggested a long-term memory advantage of threatening (i.e., angry and fearful) faces versus non-threatening faces (particularly happy and neutral faces and to a lesser extent sad faces).

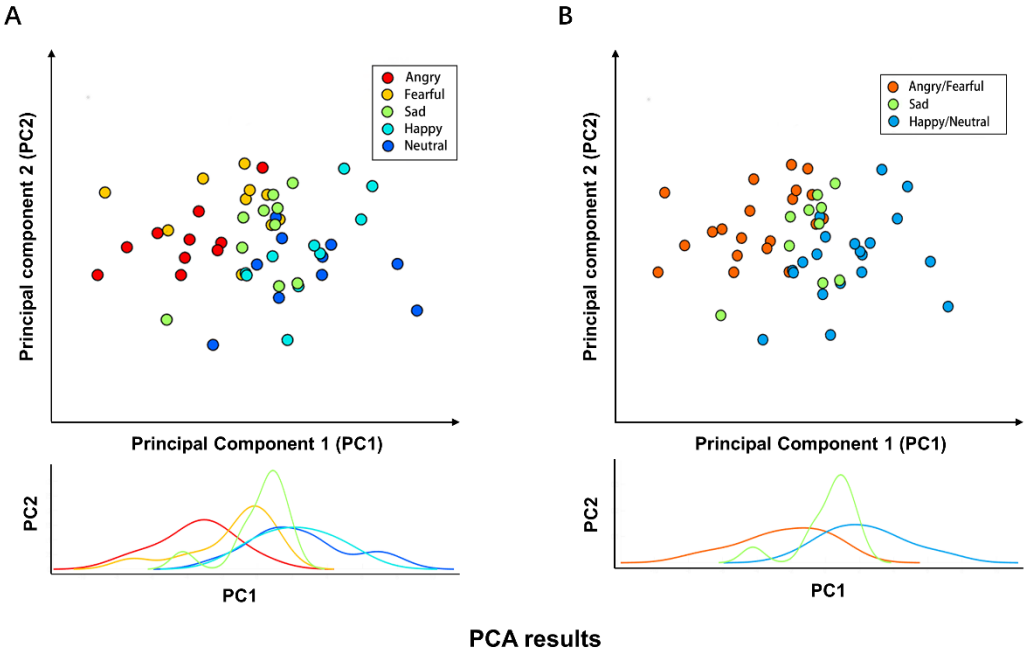

Fig. 4. PCA separation of confidence ratings for target faces in the delayed test. (*A*) Upper panel: Effects of facial expression conditions (color-coded) on memory performance along PC1 axis. Lower panel: Distribution of PC1 scores for each facial expression condition. (*B*) Upper panel:Angry and fearful faces are visualized together; neutral and happy faces are visualized together. Lower panel: Distribution of PC1 scores for each cluster.

3.2 Distinct long-term face memory patterns between subgroups

Pattern similarity analysis successfully categorized the participants into a

discriminative group (N = 44, 22 males, $rs > 0.28$, $ps < 0.05$) and a non-discriminative group (N = 39, 18 males, $rs < 0.28$, $ps > 0.05$), with the discriminative group showing better recognition for threatening faces but the non-discriminative group showing no such effect (**Supplementary Results** and **Figure S2**). The discriminative and non-discriminative group did not differ with respect to gender, age, retention interval, arousal rating for each facial expression and overall recognition performance, arguing against confounding effects of these variables on the long-term memory advantage of threatening expressions (for detailed results, see **Supplementary Results**). The distinguishable response patterns of the two subgroups were used to inform the fMRI analysis.

3.3 Neural basis during encoding

Given that long-term recognition for angry and fearful faces could be robustly separated from happy and neutral faces across the analyses, yet no clear pattern emerged for sad faces, the first-level contrast of interest was modeled as threatening (angry + fearful) > non-threatening (happy + neutral) expressions. The whole-brain group-comparison analysis revealed a significant interaction effect in the vmPFC/mOFC (two-tailed t test: k = 176; peak MNI coordinates: 27, 32, -16; $p_{\text{cluster-FWE}} < 0.05$, whole-brain corrected, **Fig. 5B**). Post-hoc analyses using extracted parameter showed that the discriminative group exhibited significantly higher reactivity towards non-threatening faces compared to threatening faces (t_{43} = -5.10, corrected $p < 0.001$, Cohen's d = 0.78) while the non-discriminative group exhibited the opposite pattern (t_{38} = 2.26, corrected $p < 0.05$, Cohen's d = 0.58) (**Fig. 5C**). Against our expectations, the examination of the

amygdala ROI did not reveal significant results.

Additionally, examining the main effects of the emotional expression and group revealed that the right middle temporal gyrus (MTG) exhibited higher reactivity towards threatening vs. non-threatening faces during encoding (k = 242; peak MNI coordinates: 60, -55, -1; $p_{\text{cluster-FWE}} < 0.05$, whole-brain corrected, see **Supplementary Figure S3**), whereas no significant main effect of group was observed, arguing against unspecific face encoding differences between the groups.

In brief, these results demonstrated that decreased vmPFC/mOFC reactivity to threatening faces during encoding might contribute to the maintenance of long-term memory for threatening faces.

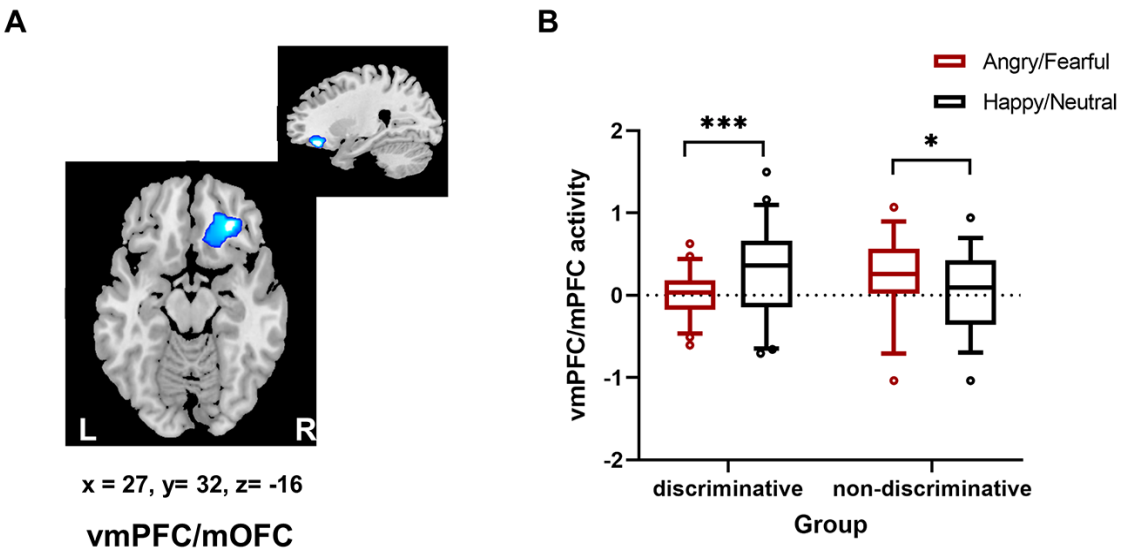

Fig. 5. The interaction effect of neural responses. (*A*) Two-sample t test of threatening (angry + fearful) vs. non-threatening (happy + neutral) contrast between discriminative group and non-discriminative group. Statistical image is displayed at $p < 0.05$ cluster-level FWE correction with a cluster-forming threshold $p < 0.001$. (*B*) Post-hoc tests on the right vmPFC/mOFC activity in the discriminative group and the non-discriminative group. *** $p < 0.001$, Holm-Bonferroni corrected.

4. Discussion

The present study systematically examined the impact of emotional face expressions during initial encounters on recognition following a retention interval of > 1.5 years. In line with our hypothesis, both univariate and multivariate analyses demonstrated that individuals better recognized threatening faces, particularly angry and fearful ones, following the long-term retention interval. The emotional expression-specific recognition advantage was not present directly after encoding and the long-term advantage of threatening faces was driven by decreased recognition of non-threatening faces over the retention period. The expression-specific face recognition pattern showed considerable inter-subject variation and a data-driven classification revealed that approximately half of the present sample demonstrated emotional expression-specific effects on face memory (discriminative group) while the other half did not (non-discriminative group). Examination of neural activation differences during encoding of threatening vs. non-threatening faces between these groups revealed that decreased vmPFC/mOFC activation to threatening vs. non-threatening faces preceded better recognition of threatening faces 1.5 years later. Together, the present findings demonstrate that threatening expressions during incidental encounters facilitated long-term face recognition and that differential encoding in the vmPFC/mOFC may contribute to expression-associated recognition differences.

Only a few previous studies investigated effects of facial expression on long-term face recognition memory using shorter retention intervals ranging from days to weeks (e.g., Anderson et al., 2006; Gupta and Srinivasan, 2009; Wang, 2013). The present

study for the first time demonstrated that recognition memory for faces with threatening emotional expressions was preserved for an extensive period of time of at least 1.5 years, consistent with findings of immediate face recognition (Grady et al., 2007; Keightley et al., 2011; Pinabiaux et al., 2013) and recognition after a 24-h delay (Wang, 2013). This finding is moreover in line with studies showing a memory advantage of negative non-facial stimuli after a retention interval of 1 year (Dolcos et al., 2005; Erk et al., 2010; Gavazzeni et al., 2012). In contrast to the expression-associated advantage at delayed recognition, no impact of facial expression on immediate recognition was observed, supporting the well-documented time-dependent emotion advantage that the preferential recognition of emotional stimuli evolved over a delay (Cahill and McGaugh, 1998; Talmi, 2013; Yonelinas and Ritchey, 2015). More importantly, our finding revealed a decrease in the recognition of non-threatening faces, which was also consistent with previous studies examining the time-dependent effects of emotion for non-facial stimuli and reporting that recognition of neutral stimuli decreased while recognition of negative stimuli remained stable (LaBar and Phelps, 1998; Sharot and Phelps, 2004). Notably, previous studies investigating emotional memory advantage emphasized the role of arousal (Bradley et al., 1992; Hamann et al., 2001; LaBar and Phelps, 1998). Indeed, arousal may bias processing toward salient information that gains processing priority and thus contribute to enhanced consolidation (Mickley et al., 2012; Ritchey et al., 2008). However, the enhanced memory for threatening faces in our study is unlikely explained by arousal based on three lines of evidence. First, despite generally lower arousal ratings for neutral faces in our study, threatening faces (angry

and fearful, respectively) did not differ from happy or sad faces with respect to arousal. Second, a moderation analysis further confirmed the lack of arousal effects by showing a non-significant interaction of facial expression category by arousal on memory performance (see **Supplementary Results**). Third, arousal ratings for each expression at encoding between discriminative and non-discriminative group were not significantly different. Given that faces are important social stimuli and threatening facial expressions such as angry and fearful signal potential danger, maintaining recognition of these stimuli over long intervals may represent an adaptive and survival-relevant mechanism (Staugaard, 2010). In contrast, faces with non-threatening expressions were forgotten over the retention interval probably due to their lower significance for future encounters (Dunsmoor et al., 2015).

An exploratory analysis further examined the brain systems that may promote the long-term recognition advantage. We observed decreased activity for threatening relative to non-threatening faces in the right vmPFC, particularly the mOFC in the discriminative group, suggesting that encoding-related activity in this region may underlie individual differences in the long-term memory advantage for threatening faces. This partly aligns with a previous study reporting an association between encoding-related activation in prefrontal regions and the amygdala with subsequently enhanced memory for negative non-facial emotional material over retention intervals of up to 1 year (Erk et al., 2010). On the other hand – and in contrast to our hypothesis - we did not observe differential coding of threatening vs. non-threatening faces in the amygdala. The vmPFC/mOFC region constitutes a core node of the network engaged

in emotional learning and value processing and, such that activity in these regions has been associated with subjective emotional experience and processing of salient emotional stimuli (Northoff, 2000; Tsukiura and Cabeza, 2011). Compared to the amygdala, the vmPFC has been suggested to play a more complex and central role in threat evaluation (Andrewes and Jenkins, 2019; Rolls, 2019) and as such is involved in both threat expression and regulation (Coccaro, et al. 2007; Davidson et al., 2000) via communication with the basolateral amygdala (Zhou et al., 2019; Delgado et al., 2016; Coccaro, et al. 2007). Moreover, the mOFC in particularly has been implicated in coding the threatening value of a stimulus in order to guide goal-directed behavior (Rudebeck and Rich, 2018). Previous neuroimaging studies showed that negative facial expressions (e.g., angry and fearful faces) elicited strong activation in OFC, particularly in mOFC (Dougherty et al., 1999; Northoff, 2000; Satterthwaite et al., 2009), and this engagement has been suggested to be related to inhibition of a behavioral response to a perceived threat (Coccaro, et al. 2007). Correspondingly, patients with OFC lesions not only judge negative facial expressions to be more approachable (Willis et al., 2010), but also display increased approach and reduced avoidance behavior for angry facial expressions (Buades-Rotger et al., 2020). Thus, a stronger differential coding of threatening versus non-threatening faces in the vmPFC/mOFC as observed in the discriminators may reflect that individual differences in threat evaluation, avoidance and regulation can impact subsequent consolidation in turn promoting the formation of stronger memory traces for threatening faces.

The findings of the present study need to be considered in the context of the

strengths and limitations of the study design. The application of data-driven multivariate PCA allowed us to detect hidden pattern in the confidence ratings in a hypothesis-free manner. The consistent results with those found in univariate analysis not only illustrate the potential of data-driven multivariate methods for analyzing behavioral responses, particularly with high inter-individual variance, but also show the robustness of the finding about long-term emotional expression effects. In addition, we employed an innovative pattern similarity analysis of behavioral response (BPSA) by capitalizing on the principal component of confidence ratings and each participant's rating pattern to unveil how subjects represent different emotional faces and identify subjects who exhibited an expression-specific recognition advantage. This new method could facilitate examination of individual differences in complex behavioral patterns, including emotional memory, and improve the description of the underlying mechanisms. On the other hand, the small number of face images for each expression condition and the low hits did not permit a trial-wise fMRI analysis comparing remembered versus non-remembered items as in the majority of previous studies examining emotional memory effects (e.g. Becker et al., 2017). Future studies using a larger set of face stimuli of each expression are needed to further explore the subsequent memory effect of emotion. Moreover, to examine the long-term memory effect with higher sensitivity and multivariate approaches, a dimensional confidence rating was used at the long-term retention interval while the immediate recognition test implemented a forced choice response. Although previous studies used a similar approach to convert dimensional confidence ratings into binary responses (Weymar et

al., 2011; Xiu et al., 2015), the results of direct comparison between immediate and delayed recognition should be interpreted with caution. Finally, our study uncovered the encoding-related neural basis of the emotional expression effects on long-term face memory. However, the memory enhancement of emotion has also been attributed to consolidation and retrieval processes (Dolcos et al., 2005; Ritchey et al., 2008; Schmidt and Saari, 2007; Sharot, et al., 2007). Investigations of neural mechanism during consolidation and retrieval stage may thus provide a more comprehensive understanding of the formation of long-term emotional face memory.

Our study provides the first evidence for a recognition advantage of threatening faces after a long-term interval of >1.5 years. Exploratory analyses further suggested that individuals who exhibited the memory advantage for threatening faces showed differential encoding of threatening versus non-threating faces in the vmPFC/mOFC as compared to individuals who did not show an emotional memory effect. These findings extend the theory of long-term emotional memory towards facial stimuli and sheds new light on the encoding-related neural basis of the preserved memory for faces with threating expressions.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Code availability

The code for the multivariate data analysis PC-corr MATLAB code is available via https://github.com/biomedical-cybernetics/PC-corr_net. Custom code that supports the findings of this study are additionally available from the corresponding author upon reasonable request.

Funding

This work was supported by the National Key Research and Development Program of China (Grant No. 2018YFA0701400), National Natural Science Foundation of China (NSFC, No 91632117, 31700998, 31530032); Fundamental Research Funds for Central Universities (ZYGX2015Z002), Science, Innovation and Technology Department of the Sichuan Province (2018JY0001).

Acknowledgments

We thank all subjects who participated in this study.

Conflict of interest

The authors declare no competing financial interests.

References

Aly, M., & Turk-Browne, N. B. (2016). Attention promotes episodic encoding by stabilizing hippocampal representations. *Proceedings of the National Academy of Sciences*, *113*(4), E420-E429.

Anderson, A. K., Yamaguchi, Y., Grabski, W., & Lacka, D. (2006). Emotional memories are not all created equal: evidence for selective memory enhancement. *Learning & Memory*, *13*(6), 711-718.

Andrewes, D. G., & Jenkins, L. M. (2019). The role of the amygdala and the ventromedial prefrontal cortex in emotional regulation: implications for post-traumatic stress disorder. *Neuropsychology review*, 1-24.

Becker, B., Steffens, M., Zhao, Z., et al. (2017). General and emotion-specific neural effects of ketamine during emotional memory formation. *Neuroimage*, *150*, 308-317.

Blanchard, D. C., & Blanchard, R. J. (1972). Innate and conditioned reactions to threat in rats with amygdaloid lesions. *Journal of comparative and physiological psychology*, *81*(2), 281.

Bowen, H. J., Kark, S. M., & Kensinger, E. A. (2018). NEVER forget: negative emotional valence enhances recapitulation. *Psychonomic Bulletin & Review*, *25*(3), 870-891.

Bradley, M. M., Greenwald, M. K., Petry, M. C., & Lang, P. J. (1992). Remembering pictures: pleasure and arousal in memory. *Journal of experimental psychology: Learning, Memory, and Cognition*, *18*(2), 379.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British journal of*

psychology, *77*(3), 305-327.

Buades-Rotger, M., Solbakk, A. K., Liebrand, M., et al. (2020). Orbitofrontal lesion patients show an implicit approach bias to angry faces. *bioRxiv*.

Cahill, L., & McGaugh, J. L. (1998). Mechanisms of emotional arousal and lasting declarative memory. *Trends in neurosciences*, *21*(7), 294-299.

Chen, L. F., & Yen, Y. S. (2007). Taiwanese facial expression image database. *Brain Mapping Laboratory, Institute of Brain Science, National Yang-Ming University, Taipei, Taiwan*.

Cohen, A. L., Soussand, L., Corrow, S. L., Martinaud, O., Barton, J. J., & Fox, M. D. (2019). Looking beyond the face area: lesion network mapping of prosopagnosia. *Brain*, *142*(12), 3975-3990.

Ciucci, S., Ge, Y., Durán, C., et al. (2017). Enlightening discriminative network functional modules behind Principal Component Analysis separation in differential-omic science studies. *Scientific reports*, *7*, 43946.

Coccaro, E. F., McCloskey, M. S., Fitzgerald, D. A., & Phan, K. L. (2007). Amygdala and orbitofrontal reactivity to social threat in individuals with impulsive aggression. *Biological psychiatry*, *62*(2), 168-178.

Darwin, C. (1872). The Expression of the Emotions in Man and Animals. London: Murray

Davidson, R. J., Jackson, D. C., & Kalin, N. H. (2000). Emotion, plasticity, context, and regulation: perspectives from affective neuroscience. *Psychological bulletin*, *126*(6), 890.

Davis, M. (1992). The role of the amygdala in fear and anxiety. *Annual review of neuroscience*, *15*(1), 353-375.

Delgado, M. R., Beer, J. S., Fellows, L. K., et al. (2016). Viewpoints: Dialogues on the functional role of the ventromedial prefrontal cortex.

Dolcos, F., LaBar, K. S., & Cabeza, R. (2005). Remembering one year later: role of the amygdala and the medial temporal lobe memory system in retrieving emotional memories. *Proceedings of the National Academy of Sciences*, *102*(7), 2626-2631.

Dougherty, D. D., Shin, L. M., Alpert, N. M., et al. (1999). Anger in healthy men: a PET study using script-driven imagery. *Biological psychiatry*, *46*(4), 466-472.

Dunsmoor, J. E., Murty, V. P., Davachi, L., & Phelps, E. A. (2015). Emotional learning selectively and retroactively strengthens memories for related events. *Nature*, *520*(7547), 345-348.

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the national academy of sciences*, *113*(28), 7900-7905.

Erk, S., Von Kalckreuth, A., & Walter, H. (2010). Neural long-term effects of emotion regulation on episodic memory processes. *Neuropsychologia*, *48*(4), 989-996.

Gavazzeni, J., Andersson, T., Bäckman, L., Wiens, S., & Fischer, H. (2012). Age, gender, and arousal in recognition of negative and neutral pictures 1 year later. *Psychology and Aging*, *27*(4), 1039.

Gong, X., Huang, Y. X., Wang, Y., & Luo, Y. J. (2011). Revision of the Chinese facial affective picture system. *Chinese mental health journal*.

Grady, C. L., Hongwanishkul, D., Keightley, M., Lee, W., & Hasher, L. (2007). The effect of age on memory for emotional faces. *Neuropsychology*, *21*(3), 371.

Gupta, R., & Srinivasan, N. (2009). Emotions help memory for faces: Role of whole and parts. Cognition and Emotion, 23, 807-816.

Hamann, S. (2001). Cognitive and neural mechanisms of emotional memory. *Trends in cognitive sciences*, *5*(9), 394-400.

Hariri, A. R., Mattay, V. S., Tessitore, A., Fera, F., & Weinberger, D. R. (2003). Neocortical modulation of the amygdala response to fearful stimuli. *Biological psychiatry*, *53*(6), 494-501.

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in cognitive sciences*, *4*(6), 223-233.

Jackson, M. C., Linden, D. E., & Raymond, J. E. (2014). Angry expressions strengthen the encoding and maintenance of face identity representations in visual working memory. *Cognition & Emotion*, *28*(2), 278-297.

Kark, S. M., & Kensinger, E. A. (2019). Post-encoding amygdala-visuosensory coupling is associated with negative memory bias in healthy young adults. *Journal of Neuroscience*, *39*(16), 3130-3143.

Keightley, M. L., Chiew, K. S., Anderson, J. A., & Grady, C. L. (2011). Neural correlates of recognition memory for emotional faces and scenes. *Social cognitive and affective neuroscience*, *6*(1), 24-37.

Kensinger, E. A., & Schacter, D. L. (2006). Amygdala activity is associated with the successful encoding of item, but not source, information for positive and negative

stimuli. *Journal of Neuroscience*, *26*(9), 2564-2570.

LaBar, K. S., & Cabeza, R. (2006). Cognitive neuroscience of emotional memory. *Nature Reviews Neuroscience*, *7*(1), 54-64.

LaBar, K. S., & Phelps, E. A. (1998). Arousal-mediated memory consolidation: Role of the medial temporal lobe in humans. *Psychological Science*, *9*(6), 490-493.

LeDoux, J. E. (1995). Emotion: Clues from the brain. *Annual review of psychology*, *46*(1), 209-235.

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., et al. (2019). Gene-Environment Interaction: A TPH2 polymorphism and early life stress shape brain architecture and anxious avoidant behavior. *bioRxiv*, 685099.

Mather, M., & Carstensen, L. L. (2003). Aging and attentional biases for emotional faces. *Psychological science*, *14*(5), 409-415.

Mickley Steinmetz, K. R., Schmidt, K., Zucker, H. R., & Kensinger, E. A. (2012). The effect of emotional arousal and retention delay on subsequent-memory effects. *Cognitive neuroscience*, *3*(3-4), 150-159.

Miendlarzewska, E. A., Ciucci, S., Cannistraci, C. V., Bavelier, D., & Schwartz, S. (2018). Reward-enhanced encoding improves relearning of forgotten associations. *Scientific reports*, *8*(1), 1-10.

Northoff, G., Richter, A., Gessner, M., et al. (2000). Functional dissociation between medial and lateral prefrontal cortical spatiotemporal activation in negative and positive emotions: a combined fMRI/MEG study. *Cerebral Cortex*, *10*(1), 93-107.

Öhman, A., Lundqvist, D., & Esteves, F. (2001). The face in the crowd revisited: a threat advantage with schematic stimuli. *Journal of personality and social psychology*, *80*(3), 381.

Pinabiaux, C., Pannier, L., Chiron, C., Rodrigo, S., Jambaqué, I., & Noulhiane, M. (2013). Memory for fearful faces across development: specialization of amygdala nuclei and medial temporal lobe structures. *Frontiers in human neuroscience*, *7*, 901.

Powell, L. J., Kosakowski, H. L., & Saxe, R. (2018). Social origins of cortical face areas. *Trends in cognitive sciences*, *22*(9), 752-763.

Ringnér, M. (2008). What is principal component analysis?. *Nature biotechnology*, *26*(3), 303-304.

Ritchey, M., Dolcos, F., & Cabeza, R. (2008). Role of amygdala connectivity in the persistence of emotional memories over time: An event-related fMRI investigation. *Cerebral Cortex*, *18*(11), 2494-2504.

Ritchey, M., LaBar, K. S., & Cabeza, R. (2011). Level of processing modulates the neural correlates of emotional memory formation. *Journal of Cognitive Neuroscience*, *23*(4), 757-771.

Rolls, E. T. (2019). The orbitofrontal cortex and emotion in health and disease, including depression. *Neuropsychologia*, *128*, 14-43.

Rudebeck, P. H., & Rich, E. L. (2018). Orbitofrontal cortex. *Current Biology*, *28*(18),

R1083-R1088.

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic bulletin & review*, *16*(2), 252-257.

Satterthwaite, T. D., Wolf, D. H., Gur, R. C., et al. (2009). Frontolimbic responses to emotional face memory: the neural correlates of first impressions. *Human brain mapping*, *30*(11), 3748-3758.

Schmidt, S. R., & Saari, B. (2007). The emotional memory effect: Differential processing or item distinctiveness? *Memory & Cognition*, *35*(8), 1905-1916.

Sharot, T., Verfaellie, M., & Yonelinas, A. P. (2007). How emotion strengthens the recollective experience: a time-dependent hippocampal process. *PLoS One*, *2*(10), e1068.

Sharot, T., & Phelps, E. A. (2004). How arousal modulates memory: Disentangling the effects of attention and retention. *Cognitive, Affective, & Behavioral Neuroscience*, *4*(3), 294-306.

Slotnick, S. D. (2017). Cluster success: fMRI inferences for spatial extent have acceptable false-positive rates. *Cognitive neuroscience*, *8*(3), 150-155.

Staugaard, S. R. (2010). Threatening faces and social anxiety: a literature review. *Clinical psychology review*, *30*(6), 669-690.

Stiernströmer, E. S., Wolgast, M., & Johansson, M. (2016). Effects of facial expression on working memory. *International Journal of Psychology*, *51*(4), 312-317.

Stretch, V., & Wixted, J. T. (1998). Decision rules for recognition memory confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and*

Cognition, *24*(6), 1397.

Talmi, D. (2013). Enhanced emotional memory: Cognitive and neural mechanisms. *Current Directions in Psychological Science*, *22*(6), 430-436.

Tardif, J., Morin Duchesne, X., Cohan, S., et al. (2019). Use of face information varies systematically from developmental prosopagnosics to super-recognizers. *Psychological Science*, *30*(2), 300-308.

Thomas, P. M., Jackson, M. C., & Raymond, J. E. (2014). A threatening face in the crowd: Effects of emotional singletons on visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(1), 253.

Tsukiura, T., & Cabeza, R. (2011). Remembering beauty: Roles of orbitofrontal and hippocampal regions in successful memory encoding of attractive faces. *Neuroimage*, *54*(1), 653-660.

Tyng, C. M., Amin, H. U., Saad, M. N., & Malik, A. S. (2017). The influences of emotion on learning and memory. *Frontiers in psychology*, *8*, 1454.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., et al. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*, *15*(1), 273-289.

Vuilleumier, P., Richardson, M. P., Armony, J. L., Driver, J., & Dolan, R. J. (2004). Distant influences of amygdala lesion on visual cortical activation during emotional face processing. *Nature neuroscience*, *7*(11), 1271-1278.

Wang, B. (2013). Facial expression influences recognition memory for faces: Robust enhancement effect of fearful expression. *Memory*, *21*(3), 301-314.

Wang, B. (2014). Effect of time delay on recognition memory for pictures: The modulatory role of emotion. *PloS one*, *9*(6), e100238.

Wang, B. (2018). Retention interval modulates the effect of negative arousing pictures on recognition memory. *Memory*, *26*(8), 1105-1116.

Weymar, M., Löw, A., & Hamm, A. O. (2011). Emotional memories are resilient to time: evidence from the parietal ERP old/new effect. *Human brain mapping*, *32*(4), 632-640.

Willis, M. L., Palermo, R., Burke, D., McGrillen, K., & Miller, L. (2010). Orbitofrontal cortex lesions result in abnormal social judgements to emotional faces. *Neuropsychologia*, *48*(7), 2182-2187.

Xu, L., Bolt, T., Nomi, J. S., et al. (2020). Inter-subject phase synchronization differentiates neural networks underlying physical pain empathy. *Social Cognitive and Affective Neuroscience*, *15*(2), 225-233.

Yonelinas, A. P., & Ritchey, M. (2015). The slow forgetting of emotional episodic memories: an emotional binding account. *Trends in cognitive sciences*, *19*(5), 259-267.

Zhou, F., Geng, Y., Xin, F., et al. (2019). Human extinction learning is accelerated by an angiotensin antagonist via ventromedial prefrontal cortex and its connections with basolateral amygdala. *Biological psychiatry*, *86*(12), 910-920.

Zhou, F., Li, J., Zhao, W., et al. (2020). Emotional contagion of pain across different social cues shares common and process-specific neural representations. *BioRxiv*.