nature neuroscience

The neuronal basis of fear generalization in humans

Selim Onat & Christian Büchel

Organisms tend to respond similarly to stimuli that are perceptually close to an event that predicts adversity, a phenomenon known as fear generalization. Greater dissimilarity yields weaker behavioral responses, forming a fear-tuning profile. The perceptual model of fear generalization assumes that behavioral fear tuning results from perceptual similarities, suggesting that brain responses should also exhibit the same fear-tuning profile. Using fMRI and a circular fear-generalization procedure, we tested this prediction. In contrast with the perceptual model, insula responses showed less generalization than behavioral responses and encoded the aversive quality of the conditioned stimulus, as shown by high pattern similarity between the conditioned stimulus and the shock. Also inconsistent with the perceptual model, object-sensitive visual areas responded to ambiguity-related outcome uncertainty. Together these results indicate that fear generalization is not passively driven by perception, but is an active process integrating threat identification and ambiguity-based uncertainty to orchestrate a flexible, adaptive fear response.

Generalization is a fundamental cognitive ability that allows organisms to deal with the complexity of real-world situations. On the basis of the similarity between the present situation and previous experiences, foraging animals may find food more efficiently. In social contexts, humans may understand others by generalizing from their own experiences or prepare defensive behaviors when encountering situations that were previously experienced to be dangerous. However, this ability must be actively controlled, and behavior should rely on similarities between events to the extent it is useful for the organism. For example, if learning is too specific, it results in unnecessary exploration of new possibilities that could have been otherwise inferred from past experience. On the other hand, when generalization is unrestrained, new opportunities with higher values might be missed or different events might be perceived as more dangerous than they really are. Thus, both extremes lead to suboptimal behavior. Nearly a century ago, Pavlov's dogs were faced with a similar problem when tested with an auditory signal that was similar to, but perceptually different from, the one that was initially associated with an appetitive outcome[1]. Under such ambiguous situations, organisms produce similar responses as the one evoked by the conditioned stimulus (CS+), leading to a phenomenon known as stimulus generalization[2–4]. In particular, when tested systematically along a well-controlled perceptual continuum, these responses smoothly decay with increasing perceptual distance and reach minimal levels for neutral, non-reinforced stimuli (CS–), thereby forming a smoothly decaying fear-tuning profile[5]. Under large boundary conditions and using only a minimal set of assumptions, these behavioral gradients can be described with relatively simple models across a large variety of sensory modalities and species[5,6].

Notably, the rate at which responses decay quantitatively characterizes the strength of the generalization[7–10]. For example, in an aversive learning context, a sharp decrease in the magnitude of responses (for example, skin conductance responses) with increasing perceptual similarity is evidence for limited fear generalization. Conversely, a response profile characterized by a slowly decaying gradient indicates increased fear generalization. Formal analysis of fear-tuning functions offers the possibility for a quantitative characterization of the generalization strength of learned aversive associations[7,11,12]. This has proven to be fruitful for the understanding of different anxiety disorders[13], such as panic disorder[14] and generalized anxiety disorder[15–17], where fear generalization is characterized by altered fear-tuning functions[13,15,17–19].

Thus, understanding how selectivity in fear generalization arises when the organism encounters a situation that is similar to the previously learned threatening stimulus is a fundamental question. The perceptual model of fear generalization[19,20] proposes that neuronal selectivity arises as a consequence of the perceptual similarity induced by the generalizing gradient. In this view, an increase in perceptual similarity between the test stimuli (that is, graded versions of CS+ and CS–) and the CS+ causes a proportionally strong positive drive to areas involved in aversive processing, which are in turn responsible for the production of fear responses[20]. Conversely increasing perceptual differences would simultaneously result in higher prefrontal inhibitory control of fear expression. In this model, the hippocampus and perceptual cortical areas have an important role and have been proposed to perform comparisons between test stimuli and a memory template of the threatening stimulus (CS+)[20]. Notably, the key prediction of the perceptual model of fear generalization is that fear selectivity of neuronal responses is directly dependent on perceptual similarity. Thus, fear tuning at the internal neuronal level would be expected to be indistinguishable from behavioral generalization profiles in terms of both their shape and selectivity. That is, neuronal fear tuning with higher selectivity than behavioral fear tuning, or responses that do not agree with the similarity structure of the generalizing stimuli (that is, monotonically decreasing similarity), would be inconsistent with this view.

Department of Systems Neuroscience, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. Correspondence should be addressed to S.O. (sonat@uke.de).

Received 29 July; accepted 14 October; published online 16 November 2015; doi:10.1038/nn.4166

However, perceptual similarity to the source of threat may not be the only relevant factor for fear generalization. Although perceptual similarity between the CS+ and other stimuli decreases along the generalization gradient by design, differential conditioning also introduces an ambiguity in the interpretation of stimuli located halfway between CS+ and CS−. At these intermediate levels, simultaneous presence of safe and threat-related stimulus features makes an unambiguous interpretation impossible (for example, safe or dangerous), thereby causing an increased unpredictability for graded stimuli with respect to different outcomes[21,22]. Hence, ambiguity-based uncertainty would be expected to increase and peak between the CS+ and CS−. Thus, in addition to perceptual similarity, the uncertainty caused by ambiguity might also contribute to the generalization of fear responses, especially given the anxiogenic role of unpredictability[23,24]. This could result in fear responses that differentiate both the CS+ and CS− from other stimuli and might therefore be important for the generation of adaptive fear responses[25,26].

An alternative view is that fear generalization is the product of an active process[12], which could deliberately lead to the generalization of fear even in the presence of perceptual discrimination of the CS+. This alternative model decouples generalization from perceptual similarity structure and predicts the existence of dissociable neuronal processes for identification of the threat on the one hand, and ambiguity-based uncertainty evaluation on the other. More precisely, this model would entail the existence of brain areas that exhibit higher selectivity than behavioral and autonomic responses (that is, hypersharp tuning) involved in the identification of the source of threat. Complementing the identification stage, neuronal processes sensitive to ambiguity-based uncertainty may be responsible for the generation of adaptive fear-responses along the generalization gradient. Thus, this model predicts that in some brain areas fear tuning may deviate both from the imposed perceptual similarity structure and the externally observed behavioral fear tuning. Furthermore, the active integration hypothesis predicts that this information (threat identification and its ambiguity) has to be integrated to jointly predict higher level threat judgments. In other words, fear selectivity might be achieved by the integration of activity profiles involved in various threat representations rather than simply emerging as a result of stimulus similarity.

To test the predictions of these two competing models, we employed a new fear-generalization procedure using face stimuli that were generated to form a perceptual similarity continuum that was circular. The circular organization allowed us to measure two-sided fear-tuning profiles and randomly assign a different pair of most dissimilar faces (that is, located at the opposite points on the similarity circle) as the CS+ and CS− for each participant. We chose faces, as they are known to activate the ventral inferotemporal cortex[27,28], to measure stimulus-evoked perception-related activation in addition to fear-related responses following an aversive conditioning procedure.

RESULTS

Autonomic and behavioral fear tuning

We collected perceptual similarity estimates for all faces using a two-alternative forced choice task and estimated the perceptual organization of faces on the basis of these binary judgments. The perceptual organization of the stimuli mirrored the circular organization that was intended (**Supplementary Fig. 1a**,**b**). Furthermore, the perceptual noise across participants was low and allowed the association of distinct perceptual states to each individual stimulus (**Supplementary Fig. 1b**). Notably, when participant's perceptual judgments were individually aligned with the selected CS+ face, similarities were almost perfectly circular: similarity decreased symmetrically and monotonically between the CS+ and CS− faces (**Supplementary Fig. 1c**).

All stimuli were shown to volunteers before and after a fear-conditioning procedure while undergoing functional magnetic resonance imaging (fMRI) scanning (for single trial structure, see **Supplementary Fig. 2a**). A baseline phase allowed us to disentangle any a priori differences that could be present between faces (**Fig. 1a**). Furthermore, to maintain comparable arousal during this baseline phase with respect to the subsequent test phase, we administered the same amount of unconditioned stimuli (UCSs, innocuous electric-shock applied on the hand). During the baseline phase, these were delivered in a fully predictable manner following the visual presentation of a shock symbol (**Fig. 1a**), thereby avoiding the association of the UCS with any face (see Online Methods for second-order balancing of the stimulus presentation order). During the conditioning phase, only the CS+ and CS− were shown, and the CS+ face predicted the occurrence of the UCS in about one-third of the trials, whereas the CS− was kept unpaired (**Fig. 1a**). The following test phase was identical to the baseline phase and the complete set of faces was again presented; however, the signboard was replaced with the CS+ face. The test phase had the same partial reinforcement rate as in the conditioning phase to avoid extinction of learned associations[20].

We characterized fear tuning by evaluating the response amplitudes to CS+, CS− and all of the intermediate faces on the basis of evoked blood oxygen level–dependent (BOLD) responses, autonomic responses (skin conductance, SCR; for the computation of SCR generalization profiles, see **Supplementary Fig. 3**) and behavioral ratings (shock expectancy ratings). CS+ trials that were paired with the UCS were not included in this analysis. During the conditioning phase, but not before, the CS+ stimulus evoked stronger autonomous responses than the CS− face (26.6% increase; pair-wise t test (28) = 3.01, P = 0.0055; **Fig. 1b**), demonstrating that successful conditioning occurred. Confirming this observation, shock-expectancy ratings following the conditioning phase (**Fig. 1c**) were also strongest for the CS+ face (t test (28) = 5.4 and 4.04 for left and right neighbors, respectively, $P < 0.0001$). During the test phase, autonomic (**Fig. 1b**), and behavioral (**Fig. 1c**) responses continued to peak on the CS+ face, and decayed smoothly with decreasing stimulus similarity, reaching minimal values for the CS−. This specific shape of fear tuning demonstrates that participants successfully learned the association between CS+ and UCS and exhibited fear generalization to perceptually similar faces, as shown previously[4,12].

Hypersharp and uncertainty-based fear tuning

We next identified brain areas exhibiting significant fear tuning using BOLD responses recorded during the test phase ($P < 0.05$, corrected). As expected, many fear-related brain areas, posterior and anterior cingulate cortices and subcollosal cortex, exhibited tuned responses (**Supplementary Table 1**). Notably, we also found a large locus in the ventro-medial prefrontal cortex (vmPFC), together with anterior insula (aIC) and hippocampus (**Fig. 2a**). We focused on these areas given their importance in fear generalization[11,15,20,29]. In addition, we also included inferotemporal cortex (ITC; **Fig. 2a**), which is important for object and face processing[27,28].

From these sites, we extracted single-subject responses evoked by each of the eight faces with the aim of better characterizing neuronal fear tuning and comparing these with behavioral and autonomic generalization profiles. Prior to conditioning, the response profiles did not exhibit any functional form and were essentially flat (**Fig. 2b**). The effect of conditioning was clearly visible as the emergence of fear

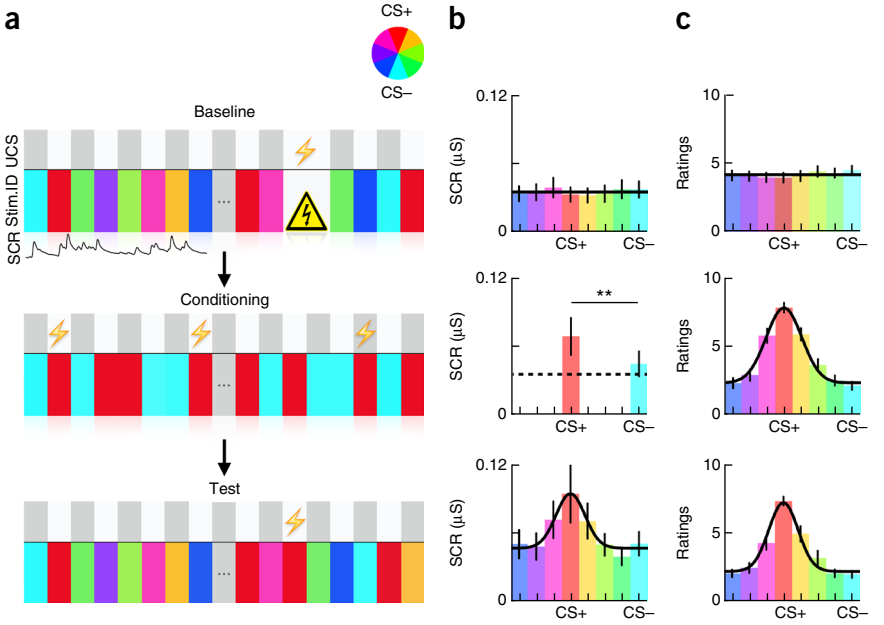

Figure 1 Experimental procedure and autonomic and behavioral generalization profiles. (**a**) During the baseline period (top), all eight faces were presented. The color wheel shows the angular position of different faces with respect to CS+ (red, CS+; cyan, CS–). A triangular signboard reliably predicted the occurrence of UCS (shock symbol). See **Supplementary Figure 2a** for details on single trial. For the conditioning phase (middle), a pair of the most dissimilar faces was randomly selected for each participant as the CS+ and CS–. Approximately one-third of CS+ trials ended together with UCS delivery (shock symbol) across conditioning and test phases. Test phase (bottom) was exactly the same as the baseline phase, but the signboard indicating the delivery of UCS was replaced by the CS+. In addition to fMRI, we continuously monitored autonomic activity with skin conductance recordings (SCR) during the presentation of the stimuli. After each experimental phase, shock expectancy ratings (that is, subjective likelihood for receiving an electric shock for different faces) allowed us to assess behavioral fear generalization. (**b**,**c**) Fear-tuning profiles based on recorded autonomic (SCR) and behavioral (ratings) responses (mean ± s.e.m.) for the three experimental phases shown in **a**. These were aligned to CS+ for each volunteer separately. Autonomic responses evoked by CS+ during the conditioning phase were ~26.6% stronger than CS– ($P = 0.0055$, t test (28) = 3.01 (indicated by **), dashed line indicates average SCR in baseline). Horizontal black lines (top) and curves (middle and bottom) show the best-fitting models, that is, null ($P > 0.5$, likelihood-ratio test) or Gaussian models ($P < 0.005$, likelihood-ratio test), respectively.

tuning during the test phase (**Fig. 2c**), mirroring the results observed at the behavioral and autonomic levels. For all areas, with the exception of ITC, fear tuning was characterized by monotonically decaying responses between the CS+ and CS– faces. Notably, different brain regions had distinctive fear tuning profiles that were characterized by differences in both the sign (for example, CS+ responses stronger than CS–, or vice versa) and width of the tuning.

We characterized the fear tuning using a Gaussian model with two parameters (α, amplitude; σ, tuning width; **Fig. 2c**) centered on the CS+ face using hierarchical Bayesian model fitting (Online Methods). The amplitude parameter can be interpreted as the differential effect between the CS+ and CS– faces after discounting for baseline shifts and is therefore a measure of tuning strength. The width parameter indicates the spread of this effect (that is, how much of the tuning strength is spread to intermediate faces) and quantifies the specificity of fear tuning.

With the exception of ITC, post-conditioning fear tuning was significantly better modeled with a Gaussian function ($P < 0.001$, likelihood-ratio test comparing null versus Gaussian models; **Fig. 2c**), whereas the null model (**Fig. 2b**) performed better before the conditioning phase, as expected. Consistent with this, amplitude parameters diverged away from zero following the conditioning phase (**Fig. 2d**). We observed that, among all tuned brain areas, aIC and a closely located locus in frontal operculum (**Supplementary Fig. 4b**) were the only regions where fear tuning was characterized by a positive amplitude parameter, indicating a relative signal increase in response to the CS+ face with respect to the CS–. In all of the remaining regions with Gaussian tuning, CS+ responses led to a deactivation, consistent with previous data from the vmPFC[17,20].

Notably, different brain regions exhibited different levels of fear selectivity (**Supplementary Fig. 4b**). The most widely tuned areas included regions located in the hippocampus and prefrontal cortex (subcollosal cortex and vmPFC). To our surprise, we found that the generalization profile recorded in the aIC was more selective than behavioral and autonomic profiles ($\sigma_{aIC} = 0.46$, $\sigma_{SCR} = 0.59$, $\sigma_{Rating(Test)} = 0.67$). The average $\sigma_{Rating(Test)}$ and σ_{SCR} values mapped onto one-sided 94% and 98.2% credibility intervals resulting from the posterior distribution of σ_{aIC}, suggesting that aIC representations were tuned more sharply than both behavioral and autonomic responses (**Fig. 2e** and **Supplementary Fig. 5a**,**b**). We also computed pair-wise differences between tuning-width parameters for each volunteer separately (**Supplementary Fig. 5c**). Parametric tests revealed that these pair-wise differences (σ_{aIC} versus σ_{SCR} and σ_{aIC} versus σ_{Rating}) were both significantly different than zero (t test (28), $P < 0.001$), corroborating hypersharp tuning of fear responses in aIC. Thus, fear tuning in aIC, despite monotonic decay, was different from other areas that had wider tuning functions. We did not observe any brain region that exhibited a wider tuning than behavioral or autonomic responses (**Fig. 2e**).

We next focused on ITC. Notably, responses in this area showed differential tuning between the intermediate test stimuli and both the CS+ and CS– (**Fig. 2c**). Thus, fear tuning was not characterized by a monotonical decrease and therefore deviated from both behavioral fear tuning and perceptual similarity structure. Formally, this resulted in a bimodal response profile, which can be described by a cosine function ($P < 0.001$, likelihood-ratio test comparing null versus cosine models; **Fig. 2c**). Notably, the cosine model explained fear tuning in ITC better than a model that categorically distinguished CS+ and CS– faces from the rest (Akaike information criterion: 2.11 versus –0.66 comparing categorical model versus cosine model, respectively). Here also, the amplitude parameter diverged from zero following the conditioning phase, but not before (**Fig. 2d**), indicating that these responses are related to fear processing. This specific fear-tuning profile observed in ITC is in agreement with an aversive representation that encodes ambiguity-based uncertainty, as responses here differentiated intermediate stimuli from both the CS+ and CS–.

Adaptive aspect of neuronal fear tuning

We reasoned that if an area is involved in the regulation of fear generalization, its tuning parameters (that is, amplitude and/or tuning width)

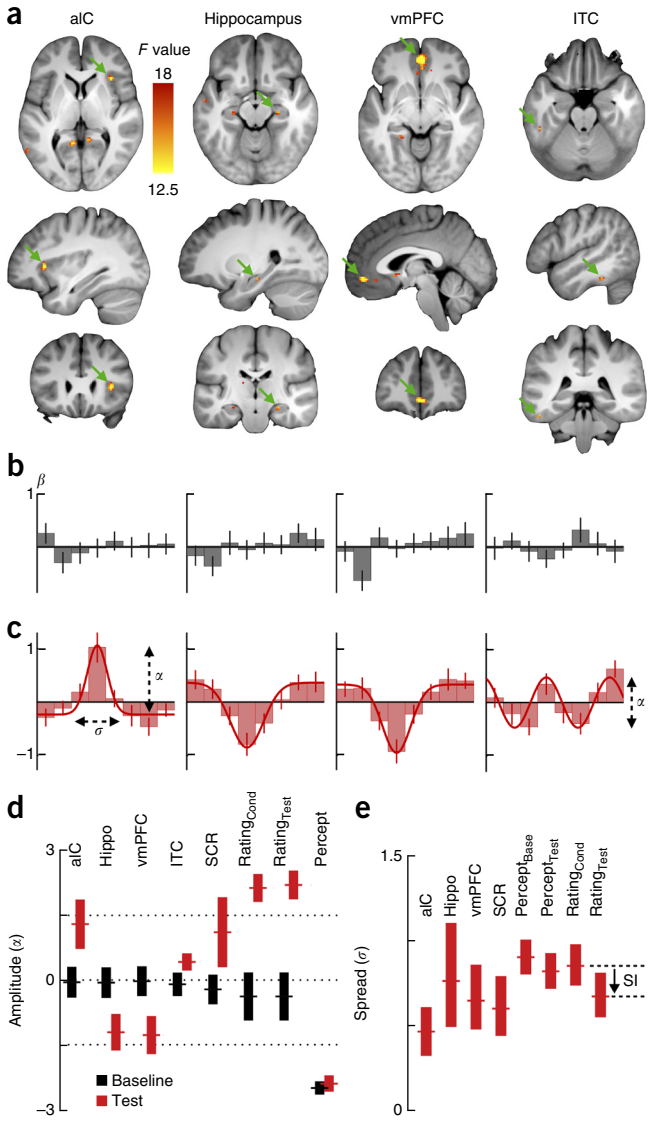

Figure 2 Brain areas exhibiting fear-tuned responses. (**a**) Thresholded statistical maps ($P < 0.05$, corrected) depict, from left to right, fear-tuned clusters in the aIC, right hippocampus, vmPFC and ITC. The functional maps are overlaid on an average anatomical image spatially normalized to MNI space. (**b**,**c**) Fear-tuning profiles of the peak-voxel in clusters depicted with green arrows in **a**, before (**b**) and after (**c**) the conditioning phase. Bars represent average evoked-responses across all volunteers for each face separately (fourth bar, CS+; eighth bar, CS−). Error bars represent s.e.m. Lines in **c** show the best fitting model to the group data (cosine model for ITC, Gaussian model for aIC, hippocampus and vmPFC, and null model for baseline). (**d**,**e**) 95% highest density interval for amplitude (**d**) and width (**e**) parameters (shown schematically in **c**) are depicted for all four regions as well as for autonomic (SCR), behavioral (rating) and perceptual tuning profiles. Horizontal lines in bars represent the mean parameter value. The spread for perceptual tuning is based on the estimates collected at the beginning ($Percept_{Base.}$) and end ($Percept_{Test}$) of the experiment (**Supplementary Fig. 1e**). $Rating_{Cond}$ and $Rating_{Test}$ indicate ratings collected before and after the test phase. Downward arrow in **e** indicates the difference in spread parameters before and after the test phase (sharpening index, SI). The tuning profile of the ITC does not have a width parameter, as it was modeled using a cosine model. Hippo, hippocampus.

should be related to behavioral fear tuning. As an indication of ongoing improvement of selectivity throughout the test phase, our data revealed that behavioral fear tuning measured before the test phase (that is, right after the conditioning phase) was substantially less tuned (that is, wider) than those measured at the end of the test phase (0.85 radians versus 0.67 radians before and after test phase, respectively; **Fig. 2e**). To quantify this increase, we devised a sharpening index (SI) comparing the tuning-width parameter before and after the test phase ($SI = \sigma_{Rating(Cond)} - \sigma_{Rating(Test)}$) for each single participant. The average SI was 0.14 radians, corresponding to an average sharpening of 16.4% (pair-wise t test (28) = 4.8, $P < 0.001$; **Fig. 2e**). This observation poses the question of how the dynamic nature of fear tuning, as shown by the SI, may result from the integration of neuronal aversive representations. We thus related this dynamic improvement in tuning selectivity to the tuning strength of previously described brain areas recorded during the test phase.

The SI was most strongly correlated with the strength of tuning (α parameter) in aIC ($r = 0.6$, [0.24, 0.79], 99.9% bootstrap confidence interval (CI); **Fig. 3a** and **Supplementary Fig. 6a**). In addition, the tuning strength in ITC was negatively correlated with SI ($r = -0.27$, [−0.7, −0.051], 99.9% CI; **Fig. 3a** and **Supplementary Fig. 6b**). We tested whether the SI could be jointly predicted by the tuning strengths of these areas, including the tuning strength of vmPFC responses ($r = -0.44$, [−0.75, −0.001], 99.9% CI; **Supplementary Fig. 6c**,**e**). A model consisting of a linear combination of tuning strengths in aIC and ITC ($SI \sim w_1\alpha_{aIC} + w_2\alpha_{ITC}$) was best amongst models that included other pair-wise combinations (that is, $SI \sim w_1\alpha_{vmPFC} + w_2\alpha_{ITC}$ or $SI \sim w_1\alpha_{vmPFC} + w_2\alpha_{aIC}$), as well as the full model (that is, $SI \sim w_1\alpha_{vmPFC} + w_2\alpha_{aIC} + w_3\alpha_{ITC}$; **Supplementary Table 2**). This indicates that tuning strength in aIC and ITC jointly explains fear selectivity observed in behavior; whereas the contribution of aIC was positive (w_1: 0.57, [0.28, 0.97] 99.5% CI), ITC had a negative contribution (w_2: −0.19, [−0.44, −0.006], 99.5% CI) to behavioral selectivity. This indicates that high aIC activity during the test phase predicts an increase in behavioral selectivity, consistent with the role of hypersharp representations in this area for threat identification. On the other hand ITC tuning had the opposite effect; a stronger tuning strength in ITC was associated with wider behavioral tuning, consistent with our interpretation of ambiguity representation in this area, resulting in wider behavioral tuning when the uncertainty associated to intermediate faces is not resolved.

In vmPFC, apart from the tuning strength, the tuning-specificity parameter also tracked behavioral selectivity as measured with the SI ($r = -0.43$, [−0.66, −0.04], 99.9% CI; **Fig. 3b** and **Supplementary Fig. 6d**,**f**). Given the negative sign of tuning strength in vmPFC (that is, highest responses for CS−), these observations are compatible with vmPFC mediating a safety signal[15]. This suggests that threat representations in vmPFC are closer to behavioral output than those in aIC and ITC, suggesting that this region might be responsible for the representation of a more abstract threat information, presumably resulting from the integration of separate aversive representations[30].

Notably, the tuning specificity in vmPFC was significantly wider than that in aIC ($\sigma_{aIC} = 0.46$ versus $\sigma_{vmPFC} = 0.65$ radians, the average σ_{vmPFC} value mapped onto 97.5% credibility interval resulting from posterior distribution of σ_{aIC}; **Fig. 2e**). This suggests that fear selectivity here cannot be explained simply by responses observed in aIC alone and points to the necessity of an additional source of information to achieve a wider fear tuning. Fear tuning in ITC that differentiates intermediate faces from both the CS+ and CS− is well suited to achieve this effect. We therefore evaluated whether fear selectivity of individual subjects observed in vmPFC can be recovered from their combined responses in aIC and ITC. Consistent with a simple additive model, we summed aIC and ITC responses for each face and participant separately (after correcting for the

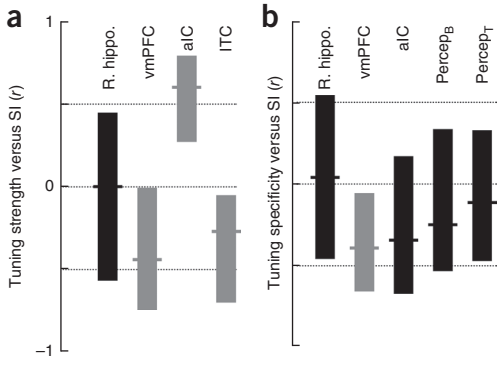

Figure 3 Correlation between SI and neuronal and perceptual tuning parameters. (**a**,**b**) Mean correlations (horizontal lines inside bars) between SI and tuning strength (**a**) and SI and tuning specificity (**b**) are shown for neuronal fear tuning, as well as perceptual tuning (last two bars in **b**). The vertical extent of the bars represents bootstrap confidence intervals (alpha = 0.01, significant correlations are shown in gray). Note that the tuning strength parameter indicates the amplitude of the Gaussian model in all comparisons except for ITC, where it represents the tuning strength of the cosine model. (R. hippo., right hippocampus; Percept$_B$ and Percept$_T$, perceptual tuning before baseline and after test phase, respectively).

sign of aIC responses, as both sources had opposite effects on the behavior) and evaluated the ensuing tuning specificity on an individual basis. We found a significant correlation ($r = 0.33$, [0.003, 0.52], 99.5% confidence intervals) between the individual tuning specificity that resulted from this integrative scheme ($\sigma_{aIC+ITC}$) and the one observed in vmPFC (σ_{vmPFC}). A linear-regression analysis based on these two variables ($\sigma_{vmPFC} \sim w_1\sigma_{aIC} + ITC$) using a robust estimation model excluded the possibility that this correlation was a result of a few outliers ($w_1 = 0.39 \pm 0.12$, $P < 0.005$). Furthermore, we tested whether the resulting fear selectivity was comparable to that in vmPFC. A pair-wise comparison of tuning-specificity parameters between vmPFC and those that resulted from this integrative scheme ($\sigma_{aIC+ITC} = 0.66$) showed no significant difference (t test (28) = –1.66, $P = 0.1$), meaning that fear selectivity arising from modeled responses combining aIC and ITC fear-tuning profiles was not distinguishable from vmPFC. These results support our view that higher level fear-selectivity, as observed in vmPFC can be recovered from the integration of hypersharp responses in aIC and bimodal response profile in ITC, thereby providing a mechanism for mapping the strength of different aversive representations onto fear-selectivity.

Nature of the hypersharp representation in insula

Previous studies[17,20] have observed fear-tuned responses in the insula in the context of fear generalization. However, univariate analyses are not well suited to investigate the content of neuronal representations[31]. Thus, we employed representational similarity analysis[32] to investigate the information represented in the aIC (**Fig. 4a**). If fear tuning, observed at the single-voxel level, relates to aversive processing, then neuronal states evoked by the CS+ would share similarity with states evoked by the aversive stimulus as such (that is, UCS before conditioning). In this view, the insula would be responsible for enacting an aversive quality for any arbitrary stimulus by integrating co-occurring information about potential harm[33]. This could mean that the neuronal representations of the CS+ face, which acquires an aversive meaning through conditioning, might gradually resemble the UCS representation. This also implies that neuronal patterns evoked by the CS+ face would gradually diverge from the originally evoked patterns by the same face before aversive learning. We tested these predictions by measuring how the similarity of neuronal patterns evoked by different stimuli changed following aversive conditioning and how these changes compared with neuronal patterns evoked by the UCS delivery during the baseline phase. Notably, during the baseline phase, UCSs were fully predictable by the onset of a shock symbol and were not associated with any face.

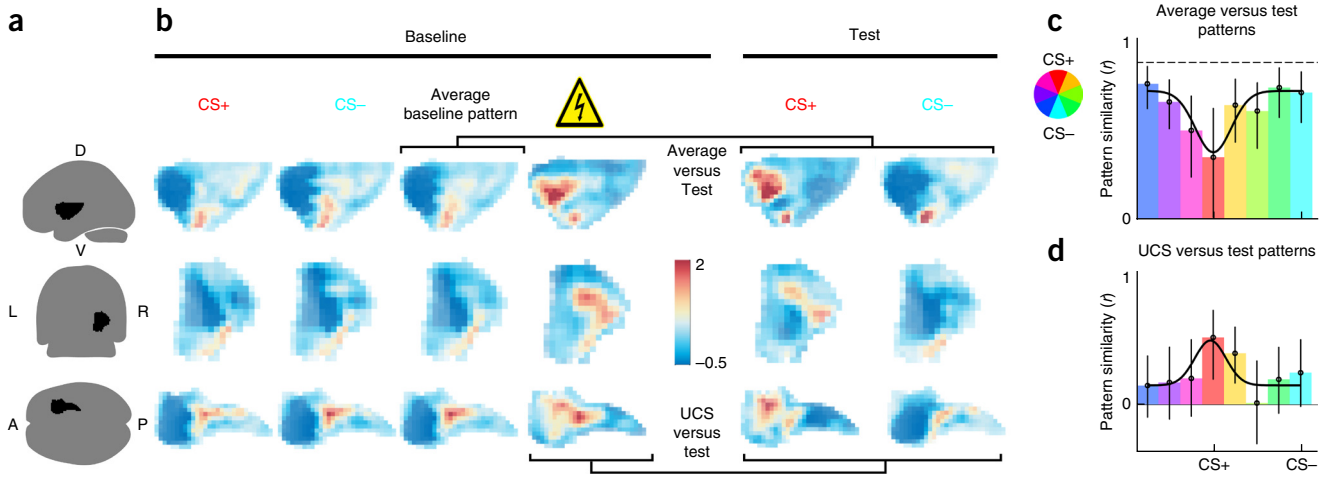

Figure 4 Effect of conditioning on insular multivariate response patterns. (**a**) Glass brain representation of the insular ROI as a binary mask. Different rows represent vertical, sagittal and horizontal views, respectively (for the exact positioning of the ROI, see **Supplementary Fig. 7a**). (**b**) Average intensity projection of evoked activity patterns in the ROI in response to CS+ and CS– before (first and second columns) and after (last two columns) the conditioning phase. The averaged pattern (third column) was computed using CS+, CS– and all of the intermediate faces (patterns shown individually in **Supplementary Fig. 7b**). The UCS-evoked pattern is depicted under the signboard symbol (fourth column). Color represents beta-weights averaged across all subjects. Common color bar for all panels. (**c**) Multivariate pattern tuning derived from the similarity measured between the average pre-conditioning pattern (third column in **b**) and the activity evoked by eight different faces recorded during test phase. Bars represent the average correlation across subjects (± 95% bootstrap confidence intervals). Dashed line depicts the mean correlation between all pairs of patterns evoked by different faces during the baseline period. (**d**) Pattern tuning resulting from similarity between the UCS-evoked pattern before the conditioning phase (fourth column in **b**) and activity patterns evoked by eight faces during the test phase. In **c** and **d**, pattern tuning was better modeled with a Gaussian function (black curve) in comparison to a null model ($P < 0.005$, likelihood-ratio test).

During the baseline phase, all face stimuli evoked similar patterns in the insula (**Fig. 4b** and **Supplementary Fig. 7b**). The average correlation between all pairs of faces was high ($r = 0.87$; **Fig. 4c**), suggesting that a single spatial pattern is sufficient to explain the neuronal dynamics by all faces before conditioning. We compared patterns evoked by all faces after conditioning to this averaged baseline pattern (**Fig. 4b**). Aversive conditioning led to a decrease in pattern similarity: the largest difference was centered on the CS+ face (**Fig. 4c**), whereas the representation of the CS− face did not exhibit any major change in comparison to the baseline phase (**Fig. 4c**). This effect was graded for intermediate (that is, between CS+ and CS−) faces, resulting in a pattern tuning that was again well-described by a Gaussian model ($P < 0.05$, likelihood-ratio test). This suggests that amplitude changes observed in the previous univariate analysis were accompanied by the gradual emergence of a novel pattern for the CS+ stimuli.

To investigate the nature of the newly emerging pattern, we tested the hypothesis that the post-conditioning pattern evoked by the CS+ might bear similarities to the pattern evoked by the UCS (recorded during the baseline period). Our data clearly show that the neuronal pattern evoked by the CS+ became more similar to the UCS pattern after conditioning (**Fig. 4d**). Again, this change in similarity was better modeled by a Gaussian than a null model ($P < 0.005$, likelihood-ratio test; **Fig. 4d**). We also investigated whether this increased similarity between the patterns evoked by the CS+ face and the UCS was also present in other areas. The increased similarity was specifically present in the aIC, but not in other areas (**Supplementary Fig. 8**). Thus, our results indicate that patterns evoked by the CS+ in the insula change with fear conditioning and become similar to the pattern evoked by the UCS. This provides evidence that hypersharp representations observed at the single-voxel level are related to the processing of aversive information.

DISCUSSION

We investigated generalization of behavioral, autonomic and neuronal fear responses following aversive learning. We characterized fear tuning on the basis of facial stimuli that were perceptually organized to form a well-controlled circular similarity continuum. As expected, fear responses observed at behavioral and autonomic levels were maximal for the CS+ and decreased smoothly with increasing perceptual distances. Analysis of BOLD responses showed that the fear-tuning parameters in many areas, most notably in hippocampus and higher cortical areas, such as vmPFC, bore a strong resemblance to behavioral fear tuning, both in their functional forms and tuning width. In contrast, we observed distinct forms of neuronal fear tuning in the aIC and the ventral ITC. In aIC, fear tuning was characterized by a hypersharp aversive representation, which was notably more selective than subjects' behavioral and autonomic reports. The information content of this hypersharp representation was in agreement with aversive processing, as shown by increased pattern similarity between the CS+ and UCS. Thus, the pattern evoked by the aversive stimulus was 'transferred' to the CS+ face following conditioning. At the same time, responses in ITC, known to be implied in perceptual processing, exhibited fear responses that were compatible with an ambiguity-based uncertainty tuning, differentiating the intermediate faces from both the CS+ and CS−. Notably, the amplitude of fear tuning in these two areas had opposite effects on behavioral selectivity on an individual basis. Although aIC tuning amplitudes predicted an increased selectivity in behavioral responses, the opposite was true for ITC tuning, consistent with an ambiguity-based uncertainty coding. Integration of these two sources of information jointly predicted the dynamic sharpening of behavioral responses and tuning-specificity in vmPFC, providing evidence that behavioral fear tuning can be understood as resulting from the integration of hypersharp responses involved in threat identification with ambiguity-based uncertainty tuning in ITC. Thus, our results provide the basis for a neuronal mechanism for the emergence of fear-selectivity based on downstream integration of activity related to threat identification and its ambiguity.

The perceptual model of fear generalization assumes a direct link between the neuronal representations of perceptual similarities and fear tuning observed at the behavioral level[20]. This is thought to provide the necessary drive to the fear network, responsible in turn for the production of fear responses and its generalization[20]. In this context, monotonically decreasing fear responses at the behavioral level are thus viewed as solely resulting from the perceptual similarity between the CS+ and other stimuli. However, we found a rich diversity of tuning profiles at different neuronal sites, indicating that the system is able to extract other variables of interest besides perceptual similarity. Supporting this view, individual perceptual performance, as assessed by visual psychophysics (**Supplementary Fig. 1e**), did not have any explanatory power for fear generalization at the behavioral level (**Fig. 3b**). Consistent with the predictions of the perceptual model, we observed that hippocampal responses were strongly tuned, but we did not find any correlation between hippocampal fear tuning and behavioral generalization.

We provide evidence that fear generalization might be based on the extraction of additional variables of interest besides perceptual similarity. In light of our results showing hypersharp and ambiguity tuning in the aIC and ITC that jointly predicted behavioral selectivity, fear generalization can be best understood as an active process. The presence of hypersharp aversive representations indicates that the system can identify rather precisely the source of threat, yet wide stimulus generalization occurs at higher level cortical areas as well as behavioral and autonomic responses. This observation suggests a mechanism that can actively widen the scope of threat to perceptually similar stimuli, rather than being passively driven by perceptual similarity. The integration of ambiguity-based uncertainty with hypersharp aversive representations is the mechanism we propose for the generation of wide fear tuning that we observed at the behavioral level. This view underlines the flexible properties of the system that are necessary for the production of adaptive fear responses.

Hypersharp tuning, as we observed, is highly relevant for the identification of the source of threat. Conceptually, this echoes the observation that the selectivity of single neurons in the auditory cortex can exceed behavioral selectivity during fear conditioning[34]. Similarly, amygdala responses have been reported to discriminate the presence of threatening stimuli better than the subjects' own reports[35]. To better understand the information represented in the insula, we turned to multivariate analyses (representational similarity analysis[32]) and investigated the information content of neuronal populations in the aIC. We observed that, following conditioning, patterns evoked by the CS+ face in the insular region became similar to the pattern evoked by the aversive stimulus as such. This observation is of particular interest because insular cortex can be considered to be a sensory cortex, responsible for enacting a rich representation of the internal homeostatic state, activated via dense bottom-up connections[36] with the rest of the body. It is therefore ideally situated for 'grounding in' different stimuli by evaluating the way how they might map onto different bodily states. In this view, the transfer of the evoked pattern by the UCS to the CS+ might be necessary for the CS+ to acquire an aversive meaning, as the CS+ predicts the same aversive outcome

and thus the same bodily state as evoked by the UCS. Notably, current evidence suggests that patients suffering from anxiety disorders exhibit signs of impaired visceral and bodily processing[37] hinting at a malfunction of the insular cortex. In support of this view and consistent with our results showing a role for the insula in threat detection, specific phobias are also characterized by abnormal insular responses[38], suggesting that selectivity in the insular region might be compromised in these patients, thereby resulting in pathological fear responses.

Fear tuning in the ITC departed from both behavioral responses and perceptual similarities. Instead of a monotonic decay between the CS+ and CS− faces, neuronal responses here differentiated both the CS+ and CS− from other faces, which is compatible with a sensitivity to the ambiguity associated with different stimuli. The precise location of the ITC locus was slightly anterior in relation to the fusiform face area. This is consistent with the observation that more abstract representations of faces, such as identities, are encoded more anteriorly with respect to the fusiform face area[39]. Responses observed in the ventral ITC were negatively correlated with behavioral fear-tuning selectivity, suggesting that reduction in uncertainty was paralleled by an increased selectivity in behavior. This is in agreement with the observation that uncertain situations can lead to higher levels of anxiety in humans and animals[40,41], resulting in an inflated threat-estimates and biased judgments[21]. These observations are relevant for anxiety disorders, which are often characterized by an impaired processing of threat uncertainty[23,42]. In this view, the resolution of uncertainty is thought to be a key mechanism with an anxiolytic effect and a necessary condition for adaptive fear responses. Furthermore, uncertainty-specific responses observed in ITC suggest that fear generalization can be adjusted to different situations in an adaptive and flexible way.

The amygdala is an important region that is typically involved in fear learning and expression of fear responses[43]. However, we did not observe a significant fear tuning in this region, even after considerably lowering our statistical threshold ($P = 0.01$). One reason might be that responses in the amygdala adapt quickly[44,45] and fear tuning might therefore be 'diluted' by averaging over the whole test phase. Repeating the same analysis for three separate temporal windows (early, middle and late) supported this, as we observed some fear tuning during the early part (**Supplementary Fig. 9**).

Future clinical studies can now characterize generalization profiles in various anxiety disorders using our methodology, and explore whether increased fear generalization stems from the absence of the precise knowledge of the source of fear (aIC) and/or a malfunction in the uncertainty evaluation system, leading both to maladaptive behavioral generalization. We can therefore predict specific contributions of different neuronal impairments on high-level threat judgments during fear learning.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank S. Schall, J. Caplan, S. Geuter, T. Kietzmann, A. Etkin, B. Knutson and K. Friston for invaluable comments, and L. Kampermann for her assistance during the data acquisition. S.O. and C.B. are supported by the DFG SFB TRR 58.

AUTHOR CONTRIBUTIONS

S.O. and C.B. conceived and designed the study. S.O. acquired the data. S.O. and C.B. analyzed and interpreted the data. S.O. drafted the manuscript. C.B. and S.O. revised the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Pavlov, I. *Conditioned Reflexes: an Investigation of the Physiological Activity of the Cerebral Cortex* (Oxford University Press Humphrey Milford, 1927).
2. Bass, M.J. & Hull, C.L. The irradiation of a tactile conditioned reflex in man. *J. Comp. Psychol.* **17**, 47–65 (1934).
3. Guttman, N. & Kalish, H.I. Discriminability and stimulus generalization. *J. Exp. Psychol.* **51**, 79–88 (1956).
4. Ghirlanda, S. & Enquist, M. A century of generalization. *Anim. Behav.* **66**, 15–36 (2003).
5. Shepard, R.N. Toward a universal law of generalization for psychological science. *Science* **237**, 1317–1323 (1987).
6. Tenenbaum, J.B. & Griffiths, T.L. Generalization, similarity, and Bayesian inference. *Behav. Brain Sci.* **24**, 629–640 (2001).
7. Lissek, S. *et al.* Generalization of conditioned fear-potentiated startle in humans: experimental validation and clinical relevance. *Behav. Res. Ther.* **46**, 678–687 (2008).
8. Resnik, J., Sobel, N. & Paz, R. Auditory aversive learning increases discrimination thresholds. *Nat. Neurosci.* **14**, 791–796 (2011).
9. Schechtman, E., Laufer, O. & Paz, R. Negative valence widens generalization of learning. *J. Neurosci.* **30**, 10460–10464 (2010).
10. Holt, D.J. *et al.* A parametric study of fear generalization to faces and non-face objects: relationship to discrimination thresholds. *Front. Hum. Neurosci.* **8**, 624 (2014).
11. Dunsmoor, J.E., Prince, S.E., Murty, V.P., Kragel, P.A. & LaBar, K.S. Neurobehavioral mechanisms of human fear generalization. *Neuroimage* **55**, 1878–1888 (2011).
12. Dunsmoor, J.E. & Murphy, G.L. Categories, concepts, and conditioning: how humans generalize fear. *Trends Cogn. Sci.* **19**, 73–77 (2015).
13. Dymond, S., Dunsmoor, J.E., Vervliet, B., Roche, B. & Hermans, D. Fear generalization in humans: systematic review and implications for anxiety disorder research. *Behav. Ther.* **26**, 561–582 (2015).
14. Lissek, S. *et al.* Overgeneralization of conditioned fear as a pathogenic marker of panic disorder. *Am. J. Psychiatry* **167**, 47–55 (2010).
15. Cha, J. *et al.* Circuit-wide structural and functional measures predict ventromedial prefrontal cortex fear generalization: implications for generalized anxiety disorder. *J. Neurosci.* **34**, 4043–4053 (2014).
16. Lissek, S. *et al.* Generalized anxiety disorder is associated with overgeneralization of classically conditioned fear. *Biol. Psychiatry* **75**, 909–915 (2014).
17. Greenberg, T., Carlson, J.M., Cha, J., Hajcak, G. & Mujica-Parodi, L.R. Ventromedial prefrontal cortex reactivity is altered in generalized anxiety disorder during fear generalization. *Depress. Anxiety* **30**, 242–250 (2013).
18. Charney, D.S., Deutch, A.Y., Krystal, J.H., Southwick, S.M. & Davis, M. Psychobiologic mechanisms of posttraumatic stress disorder. *Arch. Gen. Psychiatry* **50**, 295–305 (1993).
19. Lissek, S. Toward an account of clinical anxiety predicated on basic, neurally mapped mechanisms of Pavlovian fear-learning: the case for conditioned overgeneralization. *Depress. Anxiety* **29**, 257–263 (2012).
20. Lissek, S. *et al.* Neural substrates of classically conditioned fear-generalization in humans: a parametric fMRI study. *Soc. Cogn. Affect. Neurosci.* **9**, 1134–1142 (2014).
21. Grupe, D.W. & Nitschke, J.B. Uncertainty is associated with biased expectancies and heightened responses to aversion. *Emotion* **11**, 413–424 (2011).
22. Lake, J.I. & LaBar, K.S. Unpredictability and uncertainty in anxiety: a new direction for emotional timing research. *Front. Integr. Neurosci.* **5**, 55 (2011).
23. Grupe, D.W. & Nitschke, J.B. Uncertainty and anticipation in anxiety: an integrated neurobiological and psychological perspective. *Nat. Rev. Neurosci.* **14**, 488–501 (2013).
24. Grillon, C., Baas, J.P., Lissek, S., Smith, K. & Milstein, J. Anxious responses to predictable and unpredictable aversive events. *Behav. Neurosci.* **118**, 916–924 (2004).
25. Hsu, M., Bhatt, M., Adolphs, R., Tranel, D. & Camerer, C.F. Neural systems responding to degrees of uncertainty in human decision-making. *Science* **310**, 1680–1683 (2005).
26. Nader, K. & Balleine, B. Ambiguity and anxiety: when a glass half full is empty. *Nat. Neurosci.* **10**, 807–808 (2007).
27. Kanwisher, N., McDermott, J. & Chun, M.M. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.* **17**, 4302–4311 (1997).
28. Haxby, J.V. *et al.* Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* **293**, 2425–2430 (2001).
29. Greenberg, T., Carlson, J.M., Cha, J., Hajcak, G. & Mujica-Parodi, L.R. Neural reactivity tracks fear generalization gradients. *Biol. Psychol.* **92**, 2–8 (2013).
30. Roy, M., Shohamy, D. & Wager, T.D. Ventromedial prefrontal-subcortical systems and the generation of affective meaning. *Trends Cogn. Sci.* **16**, 147–156 (2012).
31. Dunsmoor, J.E., Kragel, P.A., Martin, A. & LaBar, K.S. Aversive learning modulates cortical representations of object categories. *Cereb. Cortex* **24**, 2859–2872 (2014).

32. Kriegeskorte, N. Representational similarity analysis – connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2** 4 10.3389/neuro.06.004.2008 (2008).
33. Visser, R.M., Scholte, H.S. & Kindt, M. Associative learning increases trial-by-trial similarity of BOLD-MRI patterns. *J. Neurosci.* **31**, 12021–12028 (2011).
34. Edeline, J.M. & Weinberger, N.M. Receptive field plasticity in the auditory cortex during frequency discrimination training: selective retuning independent of task difficulty. *Behav. Neurosci.* **107**, 82–103 (1993).
35. Whalen, P.J. *et al.* Masked presentations of emotional facial expressions modulate amygdala activity without explicit knowledge. *J. Neurosci.* **18**, 411–418 (1998).
36. Craig, A.D. How do you feel? Interoception: the sense of the physiological condition of the body. *Nat. Rev. Neurosci.* **3**, 655–666 (2002).
37. Paulus, M.P. & Stein, M.B. An insular view of anxiety. *Biol. Psychiatry* **60**, 383–387 (2006).
38. Etkin, A. & Wager, T.D. Functional neuroimaging of anxiety: a meta-analysis of emotional processing in ptsd, social anxiety disorder, and specific phobia. *Am. J. Psychiatry* **164**, 1476–1488 (2007).
39. Kriegeskorte, N., Formisano, E., Sorger, B. & Goebel, R. Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proc. Natl. Acad. Sci. USA* **104**, 20600–20605 (2007).
40. Herry, C. *et al.* Processing of temporal unpredictability in human and animal amygdala. *J. Neurosci.* **27**, 5958–5966 (2007).
41. Badia, P., Harsh, J. & Abbott, B. Choosing between predictable and unpredictable shock conditions: Data and theory. *Psychol. Bull.* **86**, 1107–1131 (1979).
42. Grillon, C. *et al.* Increased anxiety during anticipation of unpredictable aversive stimuli in posttraumatic stress disorder but not in generalized anxiety disorder. *Biol. Psychiatry* **66**, 47–53 (2009).
43. Büchel, C. & Dolan, R.J. Classical fear conditioning in functional neuroimaging. *Curr. Opin. Neurobiol.* **10**, 219–223 (2000).
44. LaBar, K.S., Gatenby, J.C., Gore, J.C., LeDoux, J.E. & Phelps, E.A. Human amygdala activation during conditioned fear acquisition and extinction: a mixed-trial fMRI study. *Neuron* **20**, 937–945 (1998).
45. Büchel, C., Morris, J., Dolan, R.J. & Friston, K.J. Brain systems mediating aversive conditioning: an event-related fMRI study. *Neuron* **20**, 947–957 (1998).

ONLINE METHODS

Participants. All participants ($n = 36$) were right-handed healthy males (age: 24.3 ± 2.9, mean ± s.d.) with no history of psychiatric or neurological illness or illegal drug use. An informed consent was obtained from all participants. They were paid about 45 euros (12 euros per h) for their participation in the experiment. We discarded seven participants who missed at least 30% of oddballs trials during the test phase, leaving a total of 29 participants. All experimental procedures were approved by the Ethics Committee of the Chamber of Physicians in Hamburg.

Stimuli. We used FaceGen software (FaceGen Modeller 2.0, Singular Inversion) to generate eight faces organized on a circular similarity continuum (**Supplementary Fig. 1a**). To this end, we initially created two facial identities at 50% gender level. We mixed these identities, while simultaneously changing their gender parameters in accord with a global circular organization. For example, the most male face (**Supplementary Fig. 1a**) was created by morphing both identities by half and making the result 100% male.

Overview of the experimental protocol. The experiment started with a perceptual test (see below) with the aim of investigating the perceptual organization of different faces (15–20 min) and proceeded with shock intensity calibration (5–10 min, see below). Participants were then placed in the scanner and passed through a familiarization phase (in the scanner, 5 min). The baseline phase took about 29 min, following this phase volunteers were taken out of the scanner for a rest period of half an hour. We next proceeded with the conditioning (14 min) and test (29 min) phases. The experiment ended with the repetition of the perceptual test.

Perceptual organization of stimuli. We used maximum likelihood difference scaling[46] (MLDS) to ensure that our stimuli were circularly organized in the perceptual domain, that is, in line with the circular positioning of faces on a two-dimensional identity-gender space. We extended this method to operate on perceptual spaces with arbitrary dimensions. At a given single trial (total 216 trials) a quadruple consisting of two pairs of faces ($F_a - F_b$ and $F_c - F_d$) is presented and participants were required to select the pair that consisted of the more similar faces. This two-alternative forced choice (2-AFC) task is preferable because it allows probing perceptual variables independently from biases and drifts associated with the decision threshold. The perceptual experiment was self-paced; however, the stimuli were turned off after 6 s and participants were required to give an answer.

These binary decisions in response to single trials are postulated to be the result of an interval comparison that operates on perceptual representations of different faces (φ_a for face a, φ_b for face b, and so on) corrupted by additive Gaussian noise, ε, with equal variance (σ_{noise}). In a given trial, the decision variable, D, is thus formulated as the difference of absolute perceptual scale differences plus additive noise, $|\varphi_a - \varphi_b| - |\varphi_c - \varphi_d| + \varepsilon$. If $D > 0$, first pair would be selected and vice versa. If two pairs are perceptually very similar to each other, D would be close to zero, and the decision would mainly depend on the additive random noise. The likelihood of selecting the first pair over the other can be expressed as $P(D > 0 \mid \varphi_i, \sigma_{noise})$. The perceptual organization of these eight faces can thus be modeled by finding those perceptual scale values that are mostly likely to generate the observed binary responses that is, that maximizes the likelihood. We used a gradient descent algorithm to find φ_i and σ_{noise} parameters that maximized the likelihood of the observed 216 choices. The optimization was initiated with eight perceptual scale values that were equally distributed along a circle. Depending on the dimensionality of the φ values, this amounted estimating 8, 16 or 24 parameters (for one-, two- or three-dimensional φ values, respectively) and 1 parameter for the noise term. The likelihood of the one-dimensional model was lowest (**Supplementary Fig. 1d**; negative log-likelihood values were high); this is most likely because one-dimensional model cannot accommodate a two-dimensional perceptual space. Supporting this view, the model where perceptual scales values were two-dimensional, performed better, and a further increment in the dimensionality improved the likelihood values only negligibly (**Supplementary Fig. 1d**). Because MLDS were separately performed for different subjects, the resulting φ_i are not necessarily aligned across different subjects. Therefore, we aligned these perceptual scale values from different volunteers to the average using Procrustes alignment. Procrustes-aligned, perceptual scale values can be seen in **Supplementary Figure 1b,c**.

With eight different stimuli, it is neither realistic nor useful to use all possible quadruplets for the MLDS experiment. We thus selected a set of quadruplets that were most informative. We ran simulations with known perceptual scale values and analyzed different methods of quadruplet generation that allowed us to best recover the hidden perceptual scale values. In the most optimal method, we selected the most optimal sequence of quadruplets for the main experiment and used this trial sequence for all runs and subjects. This allowed us to conduct the perceptual experiment within reasonable time constraints. Furthermore, the perceptual test also served as a familiarization with the stimuli before the main fMRI experiment. For the MLDS analysis shown in **Supplementary Figure 1c**, we pooled responses across the two runs that were conducted at the beginning and end of the experiment. However perceptual tuning of individual participants was separately estimated based on first and second runs of the 2-AFC task (**Supplementary Fig. 1e**).

Calibration of UCS. UCS amplitude was set to 1.5 times the subjective pain threshold that is, the current amplitude that is rated as painful on half of the trials using QUEST procedure[47] with two alternatives, "Painful" and "Not Painful", in response to UCSs with different amplitudes. To obtain a stable estimation of the pain threshold the QUEST procedure was repeated 2–3 times. Each session consisted of 16 trials with current amplitudes suggested by the QUEST algorithm. UCS consisted of a mild electric-shock generated with a direct current stimulator (Digitimer Constant Current Stimulator, Digitimer) delivered with a concentric electrode firmly connected on the on the back of the right hand. Electric shocks were trains of 5-ms pulses at 66 Hz lasting in total 100 ms. Following the calibration stage, participants were taken to the scanner without detaching the electrodes from their hands. Prior to the functional scanning, the final UCS intensity was validated to be painful but bearable (using exactly same stimulation parameters). In a few rare cases, intensity was incrementally reduced if it was judged unbearable.

fMRI recordings. Before the start of functional recordings, participants passed through a familiarization procedure of ten trials in the scanner. We monitored whether they were able to detect the oddball targets and track precisely the movement of the cross-sign (**Supplementary Fig. 2a**). Oddball stimuli consisted of a randomly selected face among the eight faces with artificially added freckles (**Supplementary Fig. 2b**). We monitored eye-movements with Eyelink 2 (SR Research). If necessary this familiarization phase was repeated a second time. Eye-movement data was monitored all along the experiment but were not recorded.

Each trial started with a fixation cross that was followed by stimulus onset after a ~700-ms delay (**Supplementary Fig. 2a**). The onset of the stimulus was triggered by the scanner pulse and lasted for 1,500 ms. Participants were required to follow the fixation-cross that jumped from the upper part (forehead region) to the lower part (mouth region) of the face at 750 ms after stimulus onset. This allowed participants to gather information about the facial identity, while at the same time removing idiosyncrasies in the exploration patterns that could potentially result in an unnecessary variance across subjects. UCS delivery occurred 100 ms before stimulus offset in the form of electric shock pulse train. Inter-trial intervals were randomly and uniformly selected to be 1–4 times the TR.

fMRI recordings were conducted before, during and after conditioning (baseline, conditioning and test phases, respectively; **Fig. 1a**). Each phase was a single block of recording. The experiment was conducted in two sessions with a 30-min break in between. During the baseline phase, we recorded responses to 8 faces (each repeated 34 times) before they were associated with any aversive outcome. UCSs (delivered ten times) were exclusively paired with an electric-shock symbol, thus avoiding aversive conditioning of any specific face stimuli at this stage. Administration of UCS during the baseline phase ensured that total number of UCSs was equated between baseline and test phases, which could otherwise potentially lead to arousal differences. In this phase, volunteers were explicitly instructed that UCSs would always occur following an electric shock symbol and never during the presentation of faces. Thus there was no uncertainty associated with different faces with respect to their outcome. Participants were required to pay attention to faces, precisely follow the fixation cross, press the response-button with their right hand as soon as the oddball target face (**Supplementary Fig. 2b**, repeated ten times) was presented and stay as immobile as possible during the recordings.

During the conditioning phase, a pair of most dissimilar faces (located on opposite sides of the circular similarity continuum), were randomly selected as the CS+ and CS– (each repeated 52 times) for each volunteer separately. No blinding was done for this assignment. The reinforcement rate was kept at ~30%. This was necessary for obtaining data in response to the presentation of CS+ face that were not affected by the responses to the UCS delivery.

The test phase was identical to the baseline phase, except that the shock symbol was replaced by the CS+ face. The rate of UCS pairing was constant across conditioning and test phases to avoid extinction. During both conditioning and test phases, participants were explicitly instructed about the occurrence of electric shocks following faces, and they were told that only one face (kept unknown) was associated with an electric shock. Furthermore the rate of oddballs and UCS delivery was kept constant in a moving window of 30 trials in all phases. This was evaluated by fitting a line to the time-course of the rate information and requiring the slope to be significantly not different from zero. This ensured that arousal was homogeneously distributed.

During all phases, we used a carry-over design[48] that balanced second-order stimulus transitions (for example, a given face is as likely to precede/follow any another face) for all stimuli making it statistically very hard to associate the occurrence of the UCS with the previous stimulus. This was done for each volunteer separately. This experimental paradigm ensured the same design efficiency for baseline and test phases. Furthermore, only those trials which were not shocked were included in the analysis, leaving the number of repetitions for all conditions identical. This was realized at the expense of presenting the CS+ face, ten additional more times during the test phase; however given the total number of stimulus presentations this imbalance is negligible. Test and baseline phases consisted of ~1,040 recorded volumes, whereas the conditioning phase consisted of ~450 volumes. Six dummy scans were discarded at the start of each phase.

At the end of each experimental phase, we gathered behavioral ratings. To this end, we presented all different faces again (repeated twice) and asked volunteers to rate the likelihood of receiving an electric shock. Participants used a visual analog scale with ten discrete values to provide their behavioral ratings of shock expectancy. The extreme points of the scale were labeled "very likely" and "not at all likely". This was conducted in the scanner using a green background color to clearly demarcate the transition from the main recording and provide a neutral context. All stimuli were presented using Psychophysics Toolbox[49] in Matlab.

Recording of autonomic signals. Electrodermal activity was measured with MRI-compatible electrodes on the palm of the left hand (thenar and hypothenar sites) connected to carbon leads (Biopac, Lead108). The signal was amplified using an analog amplifier (Biopac, MP150) and sampled at 10 Hz using CED 1401 analog-digital converter (Cambridge Electronic Design).

From the raw skin conductance recordings, we computed the phasic skin conductance drive (SCR) using a deconvolution technique[50] and used these to assess the autonomic arousal associated with the onset of individual faces at the single subject level (**Supplementary Fig. 3a**). Averaging these phasic responses separately for different conditions within a temporal window of interest (**Supplementary Fig. 3a**) we derived single subject fear-tuning profiles (**Supplementary Fig. 3b**).

fMRI acquisition and preprocessing. fMRI was performed on a 3T MR Scanner (Trio, Siemens) with a 32-channel head coil. 32 axial slices (2-mm thickness, 0% gap) were sequentially acquired using a T2*-sensitive gradient echo-planar imaging (EPI) sequence (2.02-s TR, 27-ms TE, 30° slice tilt, 90° flip angle, $2 \times 2\text{-mm}^2$ in-plane resolution, $220 \times 220 \times 64\text{-mm}^3$ field of view). Limited field of view was dorsally bounded by the ventral part of corpus collosum (**Supplementary Fig. 10**). We constrained the data analysis to those voxels that were valid across all participants. In addition, an MPRAGE structural image was acquired (voxel size $1 \times 1 \times 1$ mm^3, 240 slices) for each participant. Participants viewed the back-projected stimuli via a 45° mirror placed atop the head coil.

All functional volumes were realigned to the mean image and co-registered to anatomic images with affine transformations for each volunteer separately. Anatomical scans were DARTEL-normalized to MNI-space following gray/white matter segmentation. The same transformation was applied to co-registered functional volumes using a small smoothing kernel with 4 mm × 4 mm × 4 mm of FWHM. Anatomical locations were determined based on Harvard-Oxford Maximum Probability Atlas.

Univariate modeling of generalization profiles. Based on our hypothesis of dissociation of fear generalization into threat identification and ambiguity-based uncertainty, we aimed at detecting voxels compatible with fear tuning described either by a Gaussian or a cosine function and set up a linear regression model with parametric modulators on responses evoked by all stimuli (represented as a single vector of onsets convolved with a canonical hemodynamic response function). In a second step, the fear tuning at these selected voxels were precisely parameterized based on their responses to different faces.

Practically, identification of Gaussian-tuned neuronal areas in the context of a GLM was realized by using a set of basis functions. To account for different widths (s.d.) of Gaussian profiles, we used a Gaussian basis function in combination with a numerical approximation of the derivative of the Gaussian with respect to its standard deviation parameter ($dG/d\sigma$). The combination of these two basis functions can model a large variety of Gaussian tuning profiles. For example, wider fear tuning can be obtained by changing the weight of the second basis function. This linear approximation for fitting nonlinear functions is commonly used (for example, in SPM to account for different width or onset of the HRF). This approach makes the addition of a cosine function which best describes uncertainty tuning not necessary, as it is strongly correlated with ($dG/d\sigma$) ($r = 0.82$). In other words the $dG/d\sigma$ basis function is sensitive to cosine tuning and allowed us to avoid problems related to the orthogonalization for collinearity of regressors in the context of a GLM (that is, increased variance in the estimation[51]). We tested all combinations of these two parametric modulators ([1 −1, −1 1, 1 1, −1 −1]) to detect brain areas with fear-tuned voxels, and analyzed those that exceeded a family-wise error corrected statistical threshold of 0.05 using an F-test. This parametric model was solely used for detecting fear-tuned voxels (both Gaussian or cosine-based tuning). Notably, after identifying areas, in which signal changes could be explained by a linear combination of these basis functions, we actually wanted to precisely determine the parameters of these profiles (amplitude and standard deviation for the Gaussian and amplitude for the cosine function) and thus applied a nonlinear fitting procedure to estimate these parameters for cosine and Gaussian tuning using responses to individual faces.

The precise characterization of fear tuning at these voxels was carried out based on the single-subject beta-weights representing the activation level in response to individual faces. Using a second linear model, we thus aimed to quantify responses to different conditions. Responses to different faces were modeled by 8 separate regressors that were obtained by convolving onset times with a canonical hemodynamic response function. Two additional regressors modeled the onset of the UCS and oddball trials. Raw motion parameters (three translations and three rotations), their derivatives, squared derivatives, as well as the average time-course of the left and right ventricles were included as nuisance covariates. Beta weights representing the activation levels to individual faces were then used for the parameterization of fear-tuning profiles using a nonlinear fitting procedure (see below) as well as for multivariate analysis.

Multivariate pattern analysis. Patterns evoked by different faces before and after the aversive conditioning were computed by averaging beta-weights across all subjects. We restricted this analysis to an anatomically defined region that we obtained by merging the right insula and frontal operculum probability masks (thresholded at 10%) from the Harvard-Oxford atlas (**Fig. 3a** and **Supplementary Fig. 7a**). Together, these two regions of interest incorporated the insular site exhibiting the hypersharp tuning as well as the closely neighboring frontal opercular site. Both were the only sites that exhibited fear tuning with positive amplitude values (**Supplementary Fig. 4b**), suggesting commonalities in their function.

This analysis reflects spatial patterns at the group-level and is sensitive to activity that is common across subjects, discarding single-subject specific patterns, similar to a classical evoked analysis.

Nonlinear modeling of fear-tuning profiles. To make inferences on parameter values, we used a hierarchical Bayesian model. In the case of Gaussian model, the likelihood of observing the data given model parameters was

$$L\left[D_{p,s}(f_i)\middle|\alpha_{p,s},\sigma_{p,s},\tau_s\right] = \sum -\log\left[N\left(D_{p,s}(f_i) - G\left(\alpha_{p,s},\sigma_{p,s}\right)\middle|0,\tau_s\right)\right]$$

where $D_{p,s}(f_i)$ represents the generalization profiles for participant, p, observed in the source, s. These consist of beta-weights for 8 faces, f_{1-8}. In total there were

doi:10.1038/nn.4166

15 different sources (expectancy ratings and perceptual tunings at conditioning and test phases, electrodermal activity during test phase and BOLD responses from 10 neuronal clusters) and 29 participants. Single generalization profiles were mean-centered, taking out the non-specific component. α and σ parameters represent the amplitude and the tuning width of the zero-mean Gaussian function, $G(\alpha,\sigma)$. $G(\alpha,\sigma)$was centered on the CS+ face, therefore the location parameter was constant. With this constraint, Gaussian and Von-Mises (circular Gaussian) functions are equivalent. $N(x|0,\tau)$ is the normal probability density function with zero mean and precision parameter, τ_s. This distribution characterizes the noise levels within a given source s.

In the first level, $\alpha_{p,s}$ and $\sigma_{p,s}$ parameters modeled the amplitude and tuning-width of generalization profiles in a given source. That is, for a given source, 29 different α and σ parameters were estimated. We assumed that α parameters in a given source originate from a normal distribution with $\mu^{\alpha}_{\text{source}}$ and $\sigma^{\alpha}_{\text{source}}$ hyper-parameters (α indicates the lower level parameter). These govern the source-wide characteristics for the amplitude parameter. That is, for a given source, source-level parameters formed the prior distributions for the individual generalization profiles. Similarly, σ_p parameters were modeled as originating from a log-normal distribution with hyper-parameters $\mu^{\sigma}_{\text{source}}$ and $\sigma^{\sigma}_{\text{source}}$. A given cluster's tuning-strength and tuning-width characteristics were thus quantified with the $\mu^{\alpha}_{\text{source}}$ and $\mu^{\sigma}_{\text{source}}$ parameters, respectively. To make inference on the tuning-width of different sources, we computed highest-density intervals for the $\mu^{\sigma}_{\text{source}}$ parameter, after applying an exponential transformation. This effectively compares the median of the tuning-width parameter from different sources. For simplicity in the text we referred the $\mu^{\sigma}_{\text{source}}$ parameter as σ_{source}. For example σ_{insula} represents the tuning-width parameter of the insular region.

The dependency between all parameters is shown in **Supplementary Figure 11** using the same convention as in ref. 52. The parameter that is relevant in the context of fear generalization is $\mu^{\sigma}_{\text{source}}$, obtained during the test phase. As we had no a priori hypothesis, we assumed $\mu^{\sigma}_{\text{source}}$ to be uniformly distributed (**Supplementary Fig. 11**). Using the posterior distribution from the baseline phase as prior for the test phase would be another alternative, however these were also essentially uniformly distributed. The inference was based on the posterior distributions of these parameters and hyper-parameters obtained using Markov chain Monte Carlo (MCMC) sampling using JAGS (ver. 3.3) in a hierarchical model with two levels. As initial points for MCMC sampling we used maximum-likelihood estimators. We checked the convergence by running different runs.

For modeling uncertainty-based fear tuning, we used a cosine function, $\alpha cos(f)$, as it is symmetrical around the CS+ face. Similar to a Gaussian model, α here represents the amplitude, that is, the difference between maximum and minimum responses. f represents the number of cycles between CS+ and CS– faces. We used a strongly informative prior centered on 2 to avoid converging in local minima at different harmonic frequencies.

Statistical analysis. For fMRI analysis, we used SPM (version 8) toolbox for the statistical analysis of fMRI data ($P < 0.05$, corrected). We did not have a dedicated pilot experiment for a power analysis. However, based on realistic assumptions on effect sizes, our sample size is expected to achieve 80% power at strict statistical thresholds according to previous simulation studies[53].

The model selection between Gaussian and null models was based on likelihood-ratio test using the source hyper-parameters (see above). This effectively compared how much additional variance a given model has explained after correcting for additional number of parameters introduced in comparison to a null model, consisting of simple horizontal line.

We used parametric t test and χ^2 tests. All tests were two-sided unless otherwise stated for the statistical analysis of perceptual, behavioral and autonomic responses. Assumptions on normality required for these tests were checked using Shapiro-Wilk test. In case of the skin-conductance responses, known to have large-tailed distributions, we ensured the conclusions of parametric tests using non-parametric Wilcoxon rank sum test.

Confidence intervals for correlation and regression coefficients were computed using non-parametric tests based on bias-corrected and accelerated bootstrap method[54] with 10,000 resampling. Confidence intervals are reported within square brackets.

All regression coefficients were computed after z-transformation of variables. Akaike and Bayesian Information Criterions were computed using *fitlm* function in Matlab (MathWorks). Robust regression was conducted using *fitlm*.

A **Supplementary Methods Checklist** is available.

46. Maloney, L.T. & Yang, J.N. Maximum likelihood difference scaling. *J. Vis.* **3**, 573–585 (2003).
47. Watson, A.B. & Pelli, D.G. QUEST: a Bayesian adaptive psychometric method. *Percept. Psychophys.* **33**, 113–120 (1983).
48. Aguirre, G.K., Mattar, M.G. & Magis-Weinberg, L. de Bruijn cycles for neural decoding. *Neuroimage* **56**, 1293–1300 (2011).
49. Pelli, D.G. The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spat. Vis.* **10**, 437–442 (1997).
50. Benedek, M. & Kaernbach, C. Decomposition of skin conductance data by means of nonnegative deconvolution. *Psychophysiology* **47**, 647–658 (2010).
51. Mumford, J.A., Poline, J.-B. & Poldrack, R.A. Orthogonalization of regressors in fMRI models. *PLoS ONE* **10**, e0126255 (2015).
52. Kruschke, J.K. *Doing Bayesian Data Analysis: a Tutorial with R and BUGS* (Academic Press, 2011).
53. Desmond, J.E. & Glover, G.H. Estimating sample size in functional MRI (fMRI) neuroimaging studies: statistical power analyses. *J. Neurosci. Methods* **118**, 115–128 (2002).
54. Efron, B. & Tibshirani, R. *An Introduction to the Bootstrap* (Chapman & Hall, 1994).