Aversive Learning Changes Face-Viewing Strategies—as Revealed by Model-Based Fixation-Pattern Similarity Analysis

Lea Kampermann[1], Niklas Wilming[2], Arjen Alink[1], Christian Büchel[1], Selim Onat[1, *]

[1] Department of Systems Neuroscience, University Medical Center Hamburg-Eppendorf, Hamburg, Germany.
[2] Department of Neurophysiology and Pathophysiology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany.

* Corresponding Author(s)
Email: s.onat@uke.de

Short Title:
Similarity Analysis of Fixation Patterns during Aversive Generalization

Abstract

To avoid costly situations animals must be able to rapidly predict imminent threat based on past experience and present noisy sensory evidence. We aimed to characterize to what extent active exploration strategies, can be adaptively tuned to achieve this goal. We measured how eye-movement patterns on 8 faces, organized along a circular similarity continuum, were modified after aversive learning and generalization. Using model-based Fixation Similarity Analysis, we characterized how similarity relationships between exploration strategies were modified after volunteers learnt to pair one face (CS+) with a mild electric shock. Initially, viewing patterns reflected the circular physical similarity structure of different faces, indicating that eye-movements were guided by subtle differences between faces. Following aversive learning, the similarity structure of exploration patterns became elliptical and stretched along the adversity gradient defined by the CS+ and the most dissimilar neutral face (CS-), indicating that exploration patterns on these faces were most dissimilar. These findings suggest that the need to predict adversity introduces substantial remodelling of exploration patterns and influences entire sensorimotor loops by selecting fixation locations that can help in categorizing stimuli as aversive or safe.

Introduction

To avoid costly situations animals must be able to rapidly predict future adversity based on information actively harvested with their sensorimotor systems. However, sensory samples are noisy and the environment is complex, therefore newly encountered situations are never exactly the same as previously experienced ones. Hence, for aversive learning to be effective a careful balance between generalization and selectivity is needed [1,2]. While generalization makes it possible to promptly deploy defensive mechanisms when similar adverse situations are encountered anew [3,4], selectivity ensures that only truly aversive stimuli are recognized as aversive [5–7]. When learning is tested using stimuli that lie on a well-controlled similarity continuum, this antagonism results in smoothly decaying response amplitudes with increasing dissimilarity to the adversity predicting stimulus, thereby producing the canonical generalization profile [2,8,9]. In real-world situations however adversity predictions must be based on sensory samples collected through active exploration of information embedded in complex environments [10,11]. In humans, a central part of active exploration involves eye-movements [11–15] which can rapidly determine what information is available in a scene for recognizing adversity [16]. However, little is known about how aversive learning and its generalization influence the active exploration strategies that are fundamental to collect the necessary evidence for adversity predictions.

Here we used eye-movements as a model sensorimotor loop and compared exploration strategies before and after an aversive learning during viewing of faces. Using faces as stimuli offers an ideal test bed for investigating changes in active exploration strategies through learning. On the one hand, active viewing of faces is a key ability during daily social interactions [17,18], where detecting of minute differences in the configuration of facial elements is crucial for inferring the identity or emotional content of a face [19,20]. On the other hand, the universal spatial configuration of facial elements makes it easily possible to generate stimuli that are calibrated to form a well-controlled perceptual similarity continuum required for testing aversive generalization [7,21]. Therefore these key features make it possible to use an ecologically relevant task, while precisely controlling stimulus properties to generate a perceptual continuum to probe exploration strategies during aversive generalization. During aversive learning one randomly chosen face along this continuum (CS+) was paired with a mild electric shock (UCS) through a simple Pavlovian procedure, therefore introducing an adversity gradient based on physical similarity. This allowed us to investigate how exploration strategies deployed to collect sensory information were modified during aversive generalization.

Exploration strategies of faces are typically investigated by counting the number of fixations at predefined regions of interest [22,23]. Modifications in exploration strategies can therefore be characterized by relative changes in the number of fixations within these regions. Therefore this approach relies upon consistent changes in fixation counts across participants within the regions of interests. However, aversive learning might also influence fixation selection not by targeting different regions of the face, but by optimizing saccadic endpoints in order to provide relevant information more precisely to better predict adversity. This could lead to small shifts in exploration patterns that do not result in differences for fixation counts within predefined regions of interest. We therefore complemented count-based analyses with a variant of representational similarity analysis [24] that we term “fixation-pattern similarity analysis” (FPSA) which considers exploration patterns as multivariate entities [25,26]. FPSA assesses between-condition (dis-)similarity of eye-movement patterns for individual participants, therefore eliminating the requirement for arbitrary regions of interests (Fig. 1A). One

advantage of this approach is that it is sensitive to fine-grained changes in exploration patterns independent of whether these are accompanied with changes in fixation counts. Furthermore, FPSA has the added benefit that it can cope with idiosyncratic facial exploration strategies [25–27] as long as the similarity structure of exploration patterns changes consistently across observers.

Model-based FPSA allowed us to formulate parametric hypotheses about how aversive learning might induce changes in the similarity relationship between exploration patterns when one face along the circular continuum started to predict adversity (Fig. 1B-E, top and bottom panels). The circular similarity continuum allowed us to further decompose hypotheses about the similarity relationships onto specific and unspecific orthogonal components [28], which were centered on either the CS+/CS– or +90°/–90° faces, respectively (Fig. 1B-E, middle panels). We formulated four mutually non-exclusive hypotheses about possible changes in fixation patterns with key diagnostic values for different exploration strategies. First, if fixation selection mechanism during viewing of faces operates based on salient low-level features [10,29], we would expect exploration patterns to track the circular similarity relationships between faces (Fig. 1B). Therefore, the bottom-up saliency hypothesis predicts that a circular relationship between exploration patterns is already present before aversive learning has taken place. This would be characterized by low dissimilarity between neighboring faces (1st off-diagonal) and high dissimilarity between the opposing faces separated by 180° (4th off-diagonal, Fig. 1B left panel).

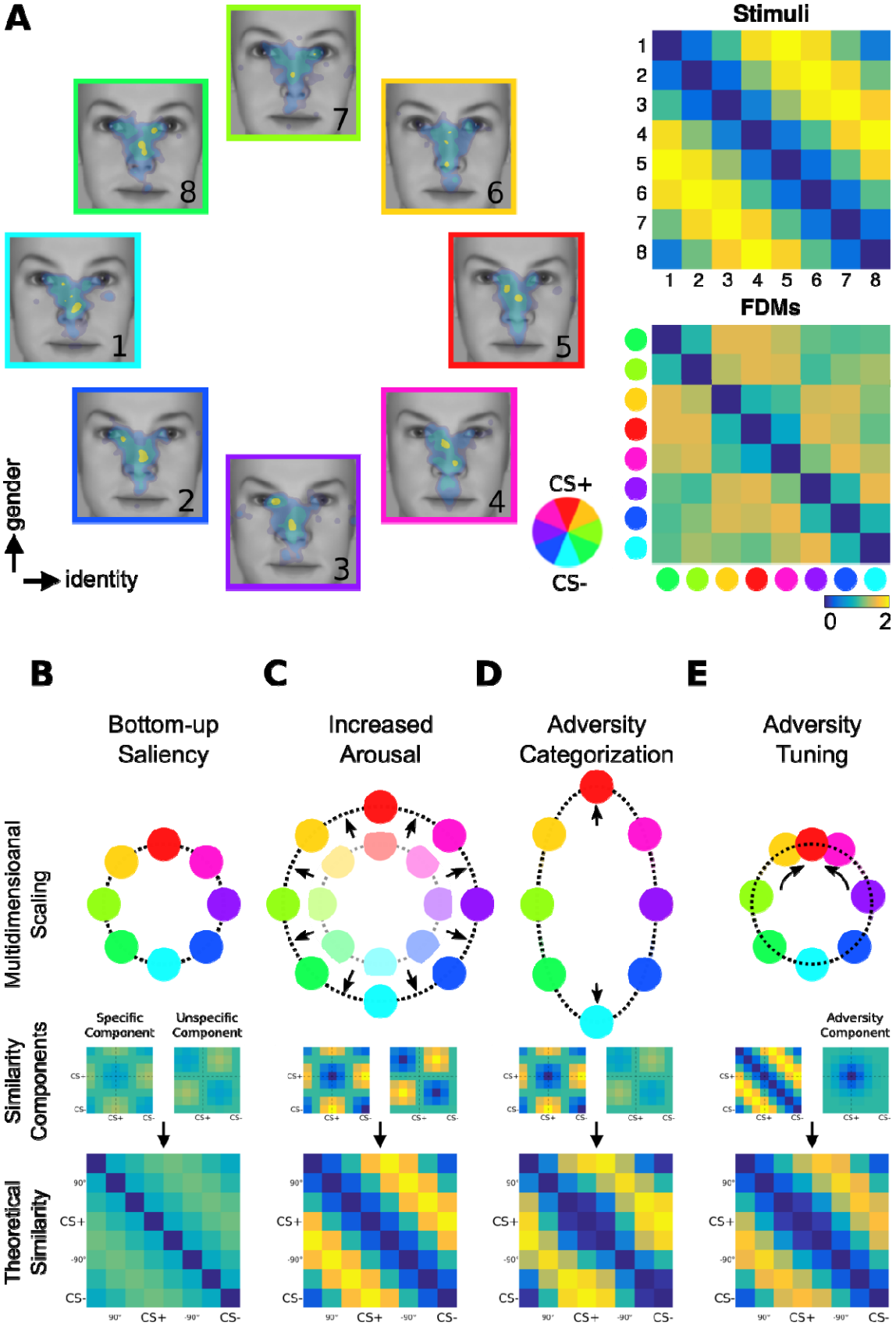

A
Stimuli
FDMs
gender
identity
CS+
CS-
0
2
B
C
D
E
Bottom-up Saliency
Increased Arousal
Adversity Categorization
Adversity Tuning
Multidimensioanal Scaling
Similarity Components
Theoretical Similarity
Specific Component
Unspecific Component
Adversity Component

Figure 1. Model-Based Fixation Similarity Analysis. (A) 8 exploration patterns (*colored frames*) overlaid on 8 face stimuli (numbered 1 to 8) calibrated to span a circular similarity continuum across two dimensions (gender and identity; see also SFig. 1 for stimuli). A pair of most dissimilar faces was randomly selected as CS+ (*red border*) and CS– (*cyan border*; see color wheel for color code). The similarity relationships among the 8 faces and the resulting exploration patterns are depicted as two 8 by 8 matrices. Physical similarity (*top right panel*) between all pair-wise combination of faces were calibrated (see methods and SFig. 2A) to have a perfect circular similarity, characterized by highest similarity (*blue*) between neighbors, and lowest similarity (*yellow*) for opposing pairs (see also SFig. 2 for calibration). Fixation Similarity Analysis summarizes the similarity relationship between the 8 exploration patterns as a symmetric 8x8 matrix (*bottom right panel*). 4th and 8th columns (and rows) are aligned with the CS+ and CS–, respectively. **(B-E)** Multidimensional scaling representation of four theoretical similarity relationships between exploration maps through learning. Each colored node represents one exploration pattern (same color scheme; *red:* CS+; *cyan:* CS–), where internode distances are proportional to dissimilarity between exploration patterns, depicted in bottom panels as 8 by 8 matrices. These matrices are further decomposed onto two orthogonal components (*middle row*), centered either on the CS+/CS– (specific component) or +90°/–90° faces (unspecific component). In **(B)**, equal contribution of components results in a circularly similar exploration patterns. In **(C)**, a stronger contribution results in a better global separation of all exploration patterns (denoted by *radial arrows* second column; *shaded nodes* depict the first hypothesis). In **(D)**, stronger contribution of the specific component results in a biased separation of exploration patterns specifically along the adversity gradient defined between the CS+ and CS– nodes. In **(E)**, a Gaussian covariance component centered on the CS+ face can specifically increase the similarity of exploration patterns for faces similar to the CS+, resulting in circularly shifted nodes (*circular arrows*) while preserving the global circularity of the similarity relationships. (same colormap as in A).

Aversive learning might lead to heightened arousal and/or an increased contribution of low-level image features on eye-movements with the objective to collect increased sensory evidence. This would result exploration strategies to mirror more closely the similarity relationships between faces (Fig. 1C) and lead to a globally increased dissimilarity between all exploration patterns. Increased arousal hypothesis therefore predicts a stronger but equal contribution of the underlying specific and unspecific components (Fig. 1C, middle panel), leading to a better separation of all exploration patterns globally. Alternatively, eye-movements may reflect a categorization process for faces as threatening or safe taking place during aversive generalization [30,31]. To achieve this, exploration strategies can be tailored to collect relevant information that jointly predicts adversity and safety. Such a fixation strategy would preferentially target locations that are discriminative of the CS+ and CS– faces. This would lead exploration patterns to become more similar for faces sharing similar features with the CS+ and CS– faces, while simultaneously predicting an increased dissimilarity between these two sets of exploration patterns (Fig. 1D, middle panel). Therefore, the 'adversity categorization' hypothesis would lead an increase of dissimilarity between exploration patterns distributed specifically along the task-relevant adversity gradient without influencing exploration patterns for intermediate faces. As a fourth possible scenario, aversive learning might result in the deployment of a new sensorimotor strategy for the adversity predicting face, thereby leading to a localized remodelling of the similarity relationships specifically around the adversity predicting CS+ face (Fig. 1E). This is supported by univariate behavioral readouts such as autonomic skin-conductance responses [21], verbal ratings of subjective adversity [32] and startle responses [33] that shows the canonical generalization profiles consisting of gradually decaying responses with increasing dissimilarity to the CS+ stimulus. Therefore, based on univariate generalization profiles, the adversity-tuning hypothesis predicts an increased similarity of exploration patterns around the CS+ face, that would decay proportionally with the increasing dissimilarity to the CS+ face.

In sum, using FPSA we analyzed the similarity relationships between exploration patterns during viewing of faces. FPSA provided us insights on how exploration strategies changed following an aversive

learning and its generalization. First, aversive learning changed exploration patterns in subtle ways that were not captured by fixation counts. Second, before learning exploration patterns showed an approximately circular similarity structure that followed the physical stimulus similarity structure. Third, after learning the similarity structure changed specifically along the adversity gradient, indicating that CS+ and CS– exploration patterns were modified, while the similarity between other faces remained largely unchanged.

Results

We created 8 face stimuli that were characterized by subtle differences in facial elements that altogether spanned a circular similarity continuum defined by two dimensions (gender and identity; see SFig. 1 for stimuli). We carefully calibrated the degree of similarity between all pairwise combinations of these faces using a biologically plausible model of the primary visual cortex [34] tuned to human contrast sensitivity (see SFig. 2 for calibration). The similarity relationship between all pair-wise faces conformed to a near perfect circular organization (Fig 1A, top right panel), such that dissimilarity varied with angular difference between faces (lowest for left and right neighbors and highest for opposing faces) with equidistant angular steps.

We confronted participants (n = 61) with this stimulus continuum before and after an aversive associative learning procedure (Fig. 2A) while measuring eye-movements during viewing of faces. During the conditioning phase, only the CS+ and CS– faces were presented and the CS+ face was partially reinforced with an aversive outcome (UCS, mild electric shock in ~30% of trials). The CS– corresponding to the face most dissimilar to the CS+ (separated by 180°) and was not reinforced. During the subsequent generalization phase, all faces were presented and the CS+ continued to be partially reinforced to prevent extinction of the previously learnt association. To ensure comparable arousal states between baseline and generalization phases, we administered UCSs also during the baseline period, however they were fully predictable as their occurrence was indicated by a shock symbol (Fig 2A).

We used autonomic skin-conductance responses (during each phase; Fig. 2B) and subjective ratings of UCS expectancy (at the end of each phase; Fig. 2C) as univariate behavioral readouts to monitor aversive learning. As expected, aversive manipulation had a profound effect on these measurements. Skin-conductance responses (SCR) recorded during the conditioning phase were on average 2.8 times higher for the CS+ face than CS– (Figure 2B, middle panel, paired t-test, $p < .001$). Explicit UCS expectancy ratings gathered at the end of the conditioning phase were also highest for the CS+ face (Figure 2C, middle panel). This showed that the CS+ face gained an aversive quality during the conditioning phase as shown by both verbal as well as autonomic recording modalities. In the subsequent generalization phase, amplitudes in both measurements decayed with increasing dissimilarity to CS+ face (Figure 2B-C, right panel) leading to a adversity-tuned profile which was well captured by a circular Gaussian curve (centered on the CS+ face, comparison to flat null model, $p < .001$, log-likelihood ratio test) in both modalities. We ruled out that aversive associations were already present before learning (comparison of flat null model and Gaussian model $p = .54$, log-likelihood ratio test; black horizontal lines in Fig. 2B-C). In summary, these univariate adversity selective measurements confirmed that learning was successfully established and transferred towards other stimuli which were perceptually similar, providing evidence for aversive generalization following learning.

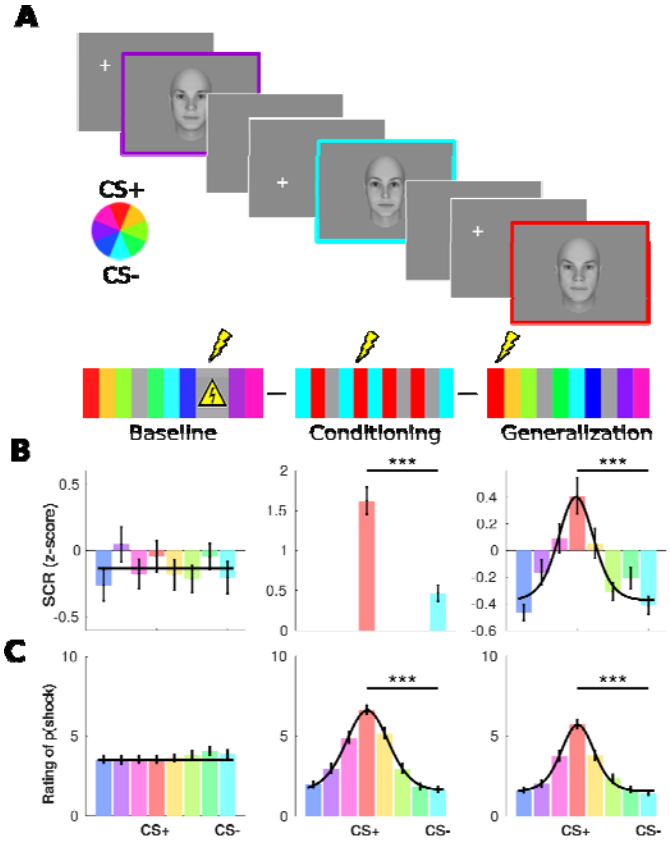


Figure 2. Univariate Characterization of Aversive Learning (A) On every trial, one out of 8 faces was presented for 1.5 seconds (conditions coded by colored frames, see color wheel). The fixation cross indicating the start of a trial was randomly placed outside of the face on either the left or right side. For each volunteer, a pair of most dissimilar faces was randomly selected as the CS+ (*red*) and CS– (*cyan*). During baseline, UCSs (indicated by shock sign) were completely predictable by a triangular signboard. During conditioning and generalization, CS+ face was paired with an aversive outcome in ~30% of trials. Null trials (*gray*) were presented resulting in a SOA of 6 or 12s. **(B)** Group-level z-scored skin-conductance responses (n = 51) and **(C)** subjective ratings of UCS expectancy (n = 61) for baseline, conditioning and generalization phases for individual faces (same color code). Responses are aligned to the CS+ for each volunteer separately. For baseline and generalization phases, the winning model (Gaussian vs. Null model) is depicted as either a black line (null model) or curve (Gaussian model). Asterisk depicts significant differences between responses to CS+ and CS– stimuli. (***: $p < .001$, *t-test*). Error bars denote SEM.

Whether eye-movements during viewing of complex stimuli such as faces also exhibit learning induced changes, and if so whether these generalize around the adversity predicting face is an open question. We first investigated this question using a fixation count-based approach. To this end, we computed fixation density maps (FDMs) for every volunteers and face separately (Fig. 3A for FDMs from two volunteers) and evaluated fixation probability within the 4 different regions of interest (left and right eyes of the face, nose and mouth [22,23]; ROIs shown in Fig. 3B as insets). We reasoned that if aversive learning has a specific influence on exploration of faces, this would result in a bell-shaped modulation of fixation counts around the CS+ face similar to SCR and verbal ratings. In line with previous report [27], left and right eyes together with the nose region were the most salient locations across the baseline and generalization phases, and attracted ~84% of all fixation density,

whereas the mouth region had only a marginal contribution with ~3% of directed density. Aversive learning increased the number of fixations directed at the nose (+4%) and mouth (+0.6%) regions at the expense of left (-3.5%) and right (-2.8%) eyes (Fig. 3B). We modeled the difference of percentages (Δ%) through learning with a categorical factor that coded for ROI identity (ROI), a second factor that indicated the absolute angular distance to the CS+ face (AngularDistance) and a third factor consisting of their interaction (Δ% ~ ROI + AngularDistance + ROI*AngularDistance). While fixation densities were significantly modulated by ROI (F = 20.18; p < .001), neither angular distances to the CS+ face (F = 0.3; p = .55) nor their interaction (F = 1.889; p = 0.12) did have any significant contribution. We complemented this linear approach with a nonlinear model, and tested how well a circular Gaussian can explain fixation counts along the similarity continuum separately for each ROIs. Model comparison on percentage changes (Fig 3B, black lines and curves) favoured the flat null model for all regions (p > 0.05, log-likelihood test), with the exception of the mouth region. Here the Gaussian model was favoured, however marginally over the null model (p = .012, log-likelihood ratio test). Therefore using the fixation count-based approach, we were able to show a specific adversity-related effect at the mouth region with increasing similarity to adversity predicting stimulus, which however exhibited only a slight increase in fixation density. However at locations that accounted for most of the fixation density were accompanied with unspecific changes that were independent of the adversity gradient.

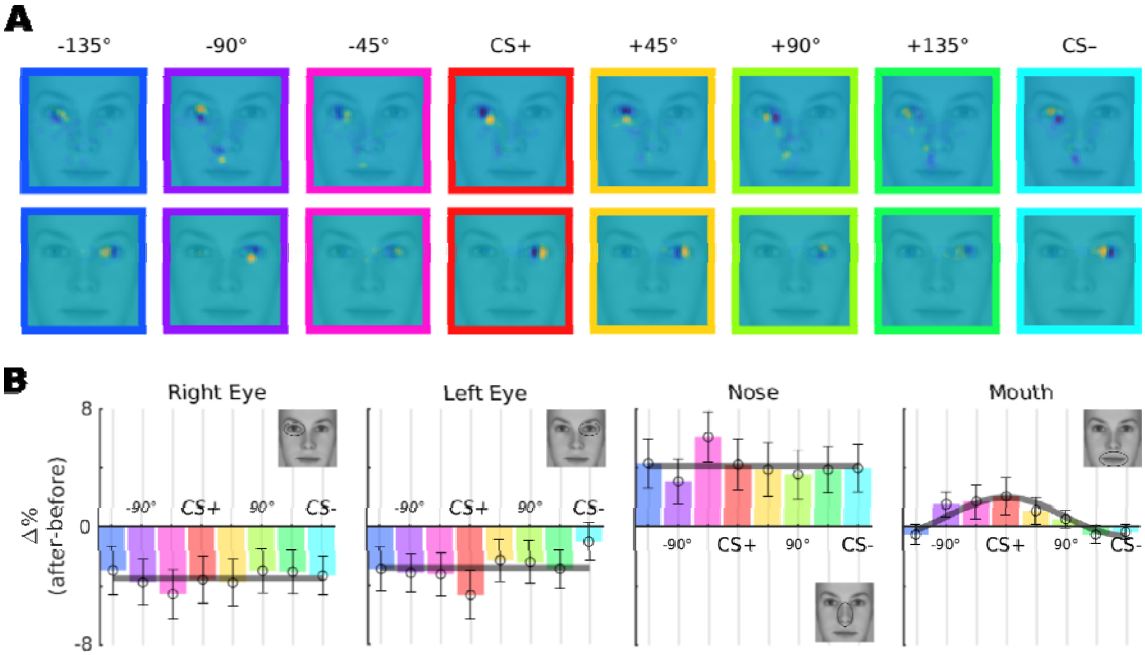


Figure 3 Impact of Aversive Learning on Fixation Counts at Four Different Regions of Interest. (A) Fixation density maps (FDMs) of two volunteers preferentially fixating on the left (*top row*) or right eye (*bottom row*) during the generalization phase. FDMs for the 8 faces are aligned to individual CS+ face (*colored frame*), and smoothed with a Gaussian kernel of 1 visual degrees. To emphasize differences between conditions, average pattern is removed from single-conditions for each volunteer separately (dark blue: less fixation than average, yellow: more fixation than average). **(B)** Percentage change in fixation density for different faces (colored bars) within 4 different regions of interests (black contours in inset). Y-axis shows the difference between generalization and baseline phases (values > 0 represent more fixation during generalization). Lines or curves indicate the winning model Gaussian model vs. Flat null model). Errorbars: SEM.

Despite the lack of a adversity specific effect at the most salient locations, a careful examination of single-subject FDMs revealed fine-grained patterning on exploration patterns that lawfully changed along the dissimilarity continuum (Fig. 3A). Notably, these differences existed within the ROIs that fixation count analysis didn't report any difference. This suggests that multivariate analysis methods [26] might therefore be more appropriate for investigating the impact of aversive learning on fixation patterns during viewing of faces. We thus investigated multivariate information content within FDMs and tested whether eye-movements deployed for the exploration of the CS+ face could be differentiated from the CS– face beyond what could be already accounted by physical differences. We evaluated how accurately a cross-validated linear classifier could discriminate FDMs on these faces before and after learning, expecting decoding accuracy to increase if aversive learning led to a differentiation of exploration patterns for CS+ vs. CS– (using a 50% holdout cross-validation with 1000 random splits of FDMs into test and training sets). The average classification performance of CS+ (59.8 ± 1.8 %; mean ± SEM across subjects) was significantly better than classification accuracy obtained before learning (52.7 ± 1.2 % classification performance; paired t-test, $p < .001$) as well as chance-level decoding (based on label permutation: 50.0 ± 0.1 %). Furthermore, classification accuracy decayed according to the typical bell-shaped curve with decreasing similarity to the CS+ face (see SFig. 3 for change in accuracy with increasing dissimilarity to CS+), suggesting a gradual deployment of the adversity specific exploration strategy with increasing similarity to the adversity predicting face. In order to understand whether this increased decoding

performance was driven from the mouth region, which previous fixation count-based analysis detected adversity-tuned density profile at the group-level, we repeated the same analysis but this time excluding the data from the mouth region. This yielded very similar results both in terms of classification accuracy and smooth decay with increasing dissimilarity, therefore excluding the possibility that decoding performance was solely driven by the adversity tuned density in the mouth region found at the group-level. Altogether these results show that exploration patterns during viewing of CS+ and CS– faces were associated with detectable differences emerging specifically with learning that could not be explained by physical differences between stimuli, or adversity-tuning present in the fixation counts. This corroborates the notion that aversive learning was associated with new exploration strategies providing strong evidence on observable modification of behavior, complementing thus SCR and verbal reports.

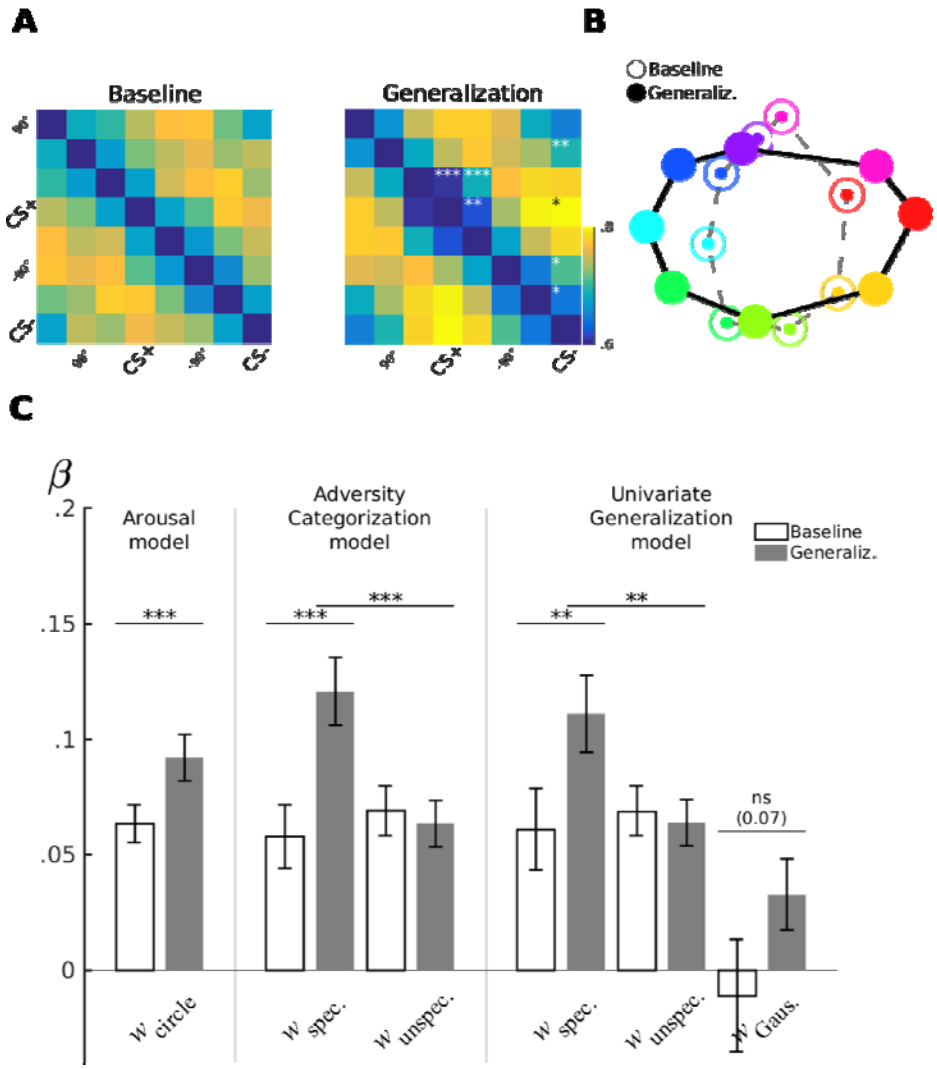


Figure 4. Multivariate analysis of Exploration Patterns and Fixation Similarity Analysis (A) Dissimilarity matrices of exploration patterns for baseline (*left panel*) and generalization phases (*right panel*). Fourth and eight columns (and rows) are aligned with each volunteer's CS+ and CS– faces, respectively. Asterisks on the upper half denote significant differences in dissimilarity values for the corresponding element between baseline and generalization phases. **(B)** Multidimensional representational similarity analysis conducted jointly on 16x16 dissimilarity matrix comprising baseline and generalization phases (not shown). Distances between nodes are proportional to the dissimilarity between corresponding FDMs (*open circles*: baseline; *filled circles*: generalization phase; same color scheme as previous). **(C)** depicts parameter estimates for the 3 tested models (*left:* bottom-up saliency; *middle:* adversity categorization; *right:* univariate generalization) fitted to individual volunteers (M ± SEM; w_{circle}: weight for the circular component, which is the sum of equally weighted specific and unspecific components; $w_{\text{specific}}/w_{\text{unspecific}}$: weight of two components centered on the CS+/CS– and +90°/-90° faces, respectively; w_{Gauss}: weight for component derived from univariate generalization profiles; **: $p < .01$; ***: $p < .001$, paired *t-test*).

However, these results cannot disentangle different hypotheses about different exploration strategies, as this requires precise characterization of all pairwise similarity relationships between exploration patterns. To gain further insights, we therefore applied a variant of representational similarity analysis [24], which we termed Fixation Similarity Analysis (FPSA). FPSA allowed us to parametrically test different hypotheses on how aversive learning could change the pairwise similarity relationships between exploration patterns, rather than the exploration patterns per se. We computed a dissimilarity matrix consisting of all pairwise comparisons of FDMs for individual volunteers (using 1 - Pearson correlation) and averaged these after separately aligning them to each volunteer's CS+ face (shown always at the 4th column and row Fig 4A). Furthermore, in order to gather an intuitive understanding learning-induced changes in the similarity geometry we used multidimensional scaling (jointly computed on the 16x16 matrices). Multidimensional scaling intuitively summarizes complicated similarity matrices by transforming observed dissimilarities as closely as possible onto distances between different nodes, therefore making it easily understandable at a descriptive level.

Already during the baseline period the dissimilarity matrix was highly structured (Fig. 4A). In agreement with a circular similarity geometry and the MDS depiction (Fig. 4B) lowest dissimilarity values (1.04 ± .01; M ± SEM) were found between FDMs of neighboring faces (i.e. first off-diagonal), whereas FDMs for faces separated by 180° exhibited significantly higher dissimilarity (1.21 ± .01 (paired t-test, $t(60) = 7.03$, $p < .001$). We investigated to what extent the observed dissimilarity structure before learning could be accounted by physical aspects of the stimulus set using the bottom-up saliency model (shown in Figure 1B). Bottom-up saliency model uses a theoretically circular similarity matrix (consisting of equally weighted sums of specific and unspecific components) as a linear predictor, therefore it estimates the global dissimilarity between exploration patterns. The circular bottom-up saliency model performed significantly better compared to a null model consisting of a constant similarity for all pairwise FDMs comparisons (for bottom-up model adjusted $r^2 = .09$; log-likelihood-ratio test for the alternative null model: $p < 10^{-5}$; $\text{BIC}_{\text{NullModel}} = -1529.3$, $\text{BIC}_{\text{BottomUp}} = -1650$; see STable 1 for the results of model fitting). We additionally fitted the bottom-up model for every volunteer separately. Model parameters at the aggregate level was significantly different than zero (Fig. 4C; $w_{\text{Circle}} = .063 \pm 0.008$, M ± SEM; $t = 7.89$, $p < 10^{-5}$) indicating that prior to learning exploration strategies mirrored the similarity structure physically present in the stimulus set. This provides evidence that fixation selection strategies are, at least to some extent, guided by physical stimulus properties.

Following aversive learning, (Figure 4A, right panel) we observed significant changes when comparing

baseline and generalization dissimilarity values element-by-element (Figure 4A, indicated by asterisks) providing evidence for learning-induced changes in the similarity relationships. The same bottom-up saliency model was again significant (adjusted r^2 = .33; $p < 10^{-5}$, log-likelihood ratio test), albeit in comparison to baseline phase performed notably better ($BIC_{BottomUp}$ = -1650 for the baseline vs. $BIC_{BottomUp}$ = -2715.5 for the generalization phase; see STable 2 for model fitting results). Critically, we found a significant increase in the model parameter from baseline to generalization phase (w_{Circle} = 0.092 ± 0.01; t = 9.13, $p < 10^{-5}$; Fig. 4C compare two leftmost bars) suggesting that dissimilarity between FDMs globally increased. Supporting this result, we found a significant interaction between the circular model and a factor that categorically differentiated pre- and post-learning periods (p = 0.009 for the interaction term). Overall, these results are compatible with the view that aversive learning led to a better separation of exploration patterns globally, in agreement with the heightened arousal model (Figure 1C), which predicted an increased contribution of the bottom-up saliency to the similarity of exploration patterns as shown by larger model parameters.

However, MDS method suggested that the pattern separation might have occurred mainly along the adversity gradient defined by the CS+ and CS– faces, whereas the separation along the orthogonal direction did not exhibit any noticeable changes (Fig.4B). We thus extended the circular bottom-up model to capture independent variance along the two different directions using the adversity categorization model (Fig. 1D). Model comparison indicated that this model performed better ($BIC_{BottomUp}$ = -2715 vs. $BIC_{AdversityCateg.}$ = -2897.3 during the generalization phase; adjusted r^2 = .44; see STable 3 and 4 for fitting results with the adversity categorization model on baseline and generalization phases, respectively). Notably this difference was accompanied by a nearly two times stronger contribution of the specific component ($w_{Specific}$ = 0.12 ± 0.014, t = 21.034; $w_{Unspecific}$ = 0.063 ± 0.01, t = 11.07; Figure 4C). Furthermore this difference was highly significant (p = 3.2 x 10^{-4}, t = -3.81, pair-wise *t-test*) indicating that the similarity relationship across the 8 faces were mainly modified along the adversity gradient. This provides strong evidence that aversive learning introduced changes in the scanning behavior specifically along the task-relevant adversity direction.

The remodelling of the similarity geometry along the adversity gradient can also be accompanied with exploration strategies that are specifically deployed for the adversity predicting face, resulting in localized changes in the similarity geometry around the CS+ face. We tested this hypothesis by augmenting the previous model with a two-dimensional Gaussian peaking on the CS+ and slowly decaying with increasing dissimilarity (Figure 1E). The model comparison procedure favored the simpler adversity categorization model over the augmented adversity tuning model ($BIC_{AdversityCateg.}$ = -2897.3 vs. $BIC_{AdversityTuning.}$= -2864.4 during the generalization phase; adjusted r^2 = .44; see STable 5, 6 for fitting results with adversity tuning model in baseline and generalization phases, respectively). Hence the increase in the number of predictor variables did not result in a significant reduction in explained variance. In line with this result, the parameter estimates for the two-dimensional Gaussian predictor were not significantly different than zero neither in baseline or generalization phases ($w_{Gaussian}$ = 0.015 ± 0.04 in baseline, p = .72, t = 0.35; $w_{Gaussian}$ = 0.07 ± 0.04 in generalization, p = 0.12, t = 1.56; Figure 4C). Furthermore, pair-wise differences between parameter estimates did not reach significance neither (p = 0.37, t = 0.89). We therefore conclude that further improvements of the adversity categorization model to include adversity-specific changes did not result in a better understanding of the adversity-induced changes in the similarity geometry of exploration strategies.

Discussion

We investigated how flexible and adaptive are active exploration strategies during generalization of an aversive learning with faces. Univariate readouts from autonomous recordings, as well as subjective verbal reports exhibited as expected the canonical generalization profiles, characterized by smoothly decaying responses with increasing dissimilarity to the CS+ face. Multivariate analyses of exploration patterns indicated that these observations were accompanied by decodable changes in the behavior, in line with an emergence of new exploration strategies with aversive learning. Furthermore, these gradually increased with the similarity to the CS+ face, resulting in adversity-tuned multivariate exploration patterns. Notably, fixation count-based approach did not detect specific effects of aversive learning on basic facial regions that were most often fixated. Using FPSA, we showed that aversive learning remodeled exploration patterns by increasing the dissimilarity specifically along the adversity gradient, while jointly merging scanning strategies for faces that were most representative for safety and threat. This was shown by the unequal contribution of orthogonal similarity components in the adversity categorization model, with the specific component having a stronger weight. Extending this model to further take into account changes in similarity relationships around the CS+ face, as predicted by the univariate bell-shaped generalization profiles, did not lead to a significant improvement. These results provide evidence for an internal cognitive process that is involved in the adversity-based categorization during aversive generalization, and argues against for the emergence of an adversity-specific exploration strategy, as predicted by univariate generalization profiles.

Humans move their eyes typically around three times per second. Hence, in comparison to autonomic responses—which inform about aggregate cognitive evaluations—and verbal evaluations such as ratings—which can interfere with the on-going task—eye-movements provide a large bandwidth of information that can be used to unveil rapidly unfolding cognitive processes [35]. However, in stark contrast with the universal organization of facial elements on a face, exploration patterns during viewing of faces are typically highly variable across individuals [25,26,36], therefore complicating their straightforward use. FPSA, in the same spirit as representational similarity analysis in fMRI [37,38], MEG [39] or EEG [40] investigates between-condition similarity of eye-movement patterns within the specific action repertoire of an individual, where averaging across volunteers would have a destructive effect. FPSA has been previously used for controlling eye-movement-related artefacts during neuronal recordings [40]. However, in this report FPSA played a central role to overcome methodological difficulties related to strong inter-individual differences. Furthermore, the key contribution of the FPSA was providing insights on how active exploration strategies were remodeled with aversive learning that could not have been predicted based on univariate generalization profiles.

Significant contribution of the bottom-up model prior to aversive learning suggests that exploration patterns reflected the physical similarity of the stimulus set. This is interesting, as it shows that fixation selection process closely tracks subtle differences in low-level local features during viewing of faces and indicates that landing points are selective for facial elements. In contrast, a holistic viewing strategy would have predicted a lack of dependence between similarity of physical features and exploration patterns. Viewing of faces can therefore be better understood by a local mechanism for fixation selection, rather than a completely holistic strategy. However, following aversive learning the bottom-up strategy was insufficient to explain observed changes in the similarity geometry. Rather, results observed here could be reconciled with a model that biases

fixation selection process with respect to facial features associated with both adversity and safety. This would effectively result in sensory evidence to be harvested from the most informative locations from stimuli along the adversity gradient. Thereby, increasing the information efficiency downstream and help the categorization process of faces as either safe or dangerous.

The effect of aversive learning on overt attention is well known. For example, fearful faces are more salient and attract more fixations than neutral ones [41–43]. Elementary visual features can benefit higher priority when they predict adversity, to the point of distracting an on-going task [16]. Most importantly, their strength gradually increases with similarity to the adversity predicting features [16]. These oculomotor saliency gradients indicate that stimuli predicting adversity benefit higher priority in the sensorimotor processing. Based on these results, one could argue that facial elements that bear resemblance to the adversity predicting face could gain higher priority and therefore would be fixated more often in a proportional manner with their similarity to the shocked face. This would normally lead to the generalization of fixation counts across facial elements that are most often fixated. However, this prediction was only partially validated at the mouth region, which was the single region where we observed an adversity-tuned fixation probability. It is therefore possible that saliency gradients observed with elementary features do not straightforwardly generalize to more complex situations involving for example faces simultaneously defined with multiple features.

The key contribution of the FPSA method was to exploit information present in pair-wise relationships between exploration patterns, and thereby achieve selectivity with respect to different hypotheses outlined here. Notably, these hypotheses could not be differentiated with univariate generalization profiles. In neuronal recordings, gradients of activity during generalization has been successfully used to characterize selectivity of aversive representations for example in different groups of subjects [44,45] or in different brain regions [7,46]. However, univariate gradients exploit only a limited extent of the available information about representations along the adversity gradient. Therefore, it will be highly informative to test these hypotheses using neuronal recordings with representational similarity analysis during the emergence of aversive representations. It is now interesting to speculate that different hypotheses we outlined here can map onto different brain systems, which are then integrated for the production of overt behavior. Therefore combined fMRI and eye-movement recordings can shed light on how the central nervous system can reorganize at the systems level following an aversive learning.

Methods

Participants

Participants were 74 naïve healthy males and females (n = 37 each) with normal (or corrected-to-normal) vision (age = 27 ± 4, M ± SD) and without history of psychiatric or neurological diseases, any medical condition or use of medication that would alter pain perception. Out of 74 participants, we discarded 13 participants who did not successfully associate the CS+ face with the UCS based on their subjective ratings at the end of the generalization phase. The exclusion criteria was based on an objective model comparison procedure testing whether observed subjective ratings could be significantly better modelled with a circular Gaussian model than a null model (see Nonlinear Modelling and Model Comparison section). We thus conducted the analyses of eye-movements on a homogenous set of participants who were aversively conditioned (n = 61, 31 males). Participants had not participated in any other study using facial stimuli in combination with aversive learning before. They were paid 12 Euros per hour for their participation in the experiment and provided written informed consent. All experimental procedures were approved by the Ethics committee of the Chamber of Physicians in Hamburg.

Stimulus Preparation and Calibration of Generalization Gradient

Using a two-step procedure, we created a final set of 8 calibrated faces (Fig. 2A) that were perceptually organized along a circular similarity continuum based on a model of the primary visual (V1) cortex. Using the FaceGen software (FaceGen Modeller 2.0, Singular Inversion, Ontario Canada) we created two gender-neutral facial identities and mixed these identities (0%/100% to 100%/0%) while simultaneously changing the gender parameters in two directions (more male or female). In the first step, we created a total of 160 faces by appropriately mixing the gender and identity parameters to form 5 concentric circles (see Supplementary Figure 1) based on FaceGen defined parameter values for gender and identity. Using a simple model of the primary visual cortex known to closely mirror human perceptual similarity judgments [34], we computed V1 representations for each face after converting them to grayscale. The spatial frequency sensitivity of the V1 model was adjusted to match human contrast sensitivity function with bandpass characteristics between 1 and 12 cycles/degree, peaking at 6 cycles/degrees [47]. The V1 model consists of pair of Gabor filters in quadrature at five different spatial scales and eight orientations. The activity of these 40 channels were averaged in order to obtain one single V1 representation per face. We characterized the similarity relationship between the V1 representations of 160 faces using multidimensional scaling analysis with 2 dimensions. As expected, while two dimensions explained a large variance, the improvement between two and three dimensions was only minor, providing thus evidence that the physical properties of the faces were indeed organized along two-dimensions (stress values for 1D, 2D and 3D resulting from the MDS analysis were 0.42, .04, .03, respectively). The transformation between the coordinates of the FaceGen software values (gender and identity mixing values) and coordinates returned by the MDS analysis allowed us to gather FaceGen coordinates that would correspond to a perfect circle in the V1 model. In the second step, we thus generated 8 faces that corresponded to a perfect circle. This procedure ensured that faces used in this study were organized perfectly along a circular similarity continuum according to a simple model of primary visual cortex with well-defined bandpass characteristics (Fig. 2B) known to mirror human similarity judgments. Furthermore it ensured that dimensions of gender and identity

introduced independent variance on the faces.

To present these stimuli we resized them to 1000x1000 pixels (originals: 400x400) using bilinear interpolation, and slightly smoothed with a Gaussian kernel of 5 pixels with full-width at half maximum of 1.4 pixels to remove any possible pixel artifacts that could potentially lead participants to identify faces. Faces were then normalized to have equal luminance and root-mean-square contrast. The gray background was set to the same luminance level ensuring equal brightness throughout of the experiment. Stimuli are available in [48].

Faces were presented on a 20" monitor (1600 x 1200 pixels, 60 Hz) using Matlab R2013a (Mathworks, Natick MA) with psychophysics toolbox [49,50]. The distance of the participants' eyes to the stimulus presentation screen was 50 cm. The center of the screen was at the same level as the participants' eyes. Faces spanned horizontally ~17° and vertically ~30°, aiming to mimic a typical face-to-face social situation.

Experimental paradigm

The fear conditioning paradigm (similar to [7]) consisted of baseline, conditioning and test (or generalization) phases (Fig. 2C-D). Four equivalent runs with exactly same number of trials were used during baseline (1 run) and generalization phases (3 runs) consisting of 120 trials per run (~10 minutes). Every run started with an eye-tracker calibration. Between the runs participants took a break and continued with the next run in a self-paced manner. We avoided having more than 1 runs in the baseline period in order not to induce fatigue in participants. In all three phases, subjects were instructed to fixate fixation crosses thoroughly and to press a button when an oddball stimulus appeared on the screen. This consisted of a blurred unrecognizable face. At each run during the baseline and generalization phases, 8 faces were repeated 11 times, UCS trials occurred 5 times and one oddball was presented. We presented 26 null trials with no face presented but otherwise the same trial structure (see below sequence optimization). In order to keep arousal levels comparable to the generalization phase, UCSs were also delivered during baseline, however they were fully predictable by a shock symbol therefore avoiding any face to UCS associations. During the conditioning phase, participants saw only the CS+ and the CS– faces among the 8 faces (and the null trials). These consisted of 2 most dissimilar faces separated by 180° on the circular similarity continuum and randomly assigned for every participant in a balanced manner. The conditioning was 124 trials long (~10 minutes) and CS+ and CS– faces were repeated 25 times. CS+ faces were additionally presented 11 times with the UCSs, resulting in a reinforcement rate of ~30 %. The same reinforcement ratio was used during the subsequent generalization phase in order to avoid extinction of the learnt associations. Participants were instructed that the delivery of UCSs during baseline would not be associated with faces, however in the following conditioning and test phases they were instructed that shocks would be delivered after particular faces have been presented.

Faces were presented using a rapid-event design with a stimulus onset asynchrony of 6 seconds and stimulus duration of 1.5 seconds. The presentation sequence was optimized using a modified m-sequence with 11 different conditions [51,52] (8 faces, UCS, oddball, null). An m-sequence is preferred as it balances all transitions from condition *n* to *m* (thus making the sequence as unpredictable as possible for the participant) while providing an optimal design efficiency (thus making deconvolution of autonomic skin conductance responses more reliable). In order to achieve the required reinforcement ratio (~30%), we randomly pruned UCS trials and transformed them to null trials. Similarly oddball trials were removed to have an overall rate of ~1%. This resulted in a total of 26 null trials. This deteriorated the efficiency of the m-sequence, however was a good

compromise as the resulting sequence was still much more efficient than a random sequence. Resulting from the intermittent null trials, SAOs were 6 or 12 seconds approximately exponentially distributed.

Face onsets were preceded by a fixation-cross which appeared randomly outside of the face either on the left or right side along an imaginary circle (r = 19.6°, +/- 15° above and below the horizontal center of the image). The side of fixation-cross was balanced across conditions. The calibrated electric shock was delivered as unconditioned stimulus (UCS) before the offset of the CS+ face.

Calibration and Delivery of Electric stimulation

Mild electric shocks were delivered by a direct current stimulator (Digitimer Constant Current Stimulator, Hertfordshire UK), applied by a concentric electrode (WASP type, Speciality Developments, Kent UK) that was firmly connected to the back of the right hand and fixated by a rubber glove to ensure constant contact with the skin. Shocks were trains of 5-ms pulses at 66Hz, with a total duration of 100 ms. The intensity of the electric shock applied during the experiment was calibrated for each participant before the start of the experiment. Participants underwent a QUEST procedure [53] presenting UCSs with varying amplitudes selected by an adaptive algorithm and were required to report whether a given trial was "painful" or "not painful" in a binary fashion using a sliding bar. The subjective pain threshold was the intensity that participants would rate as "painful" with a probability of 50%. The QUEST procedure was repeated twice to account for sensitization/habituation effects, thus obtaining a reliable estimate. Each session consisted of 12 stimuli, starting at an amplitude of 1mA. The amplitude used during the experiment was 2 times this threshold value. Before starting the actual experiment, participants were asked to confirm whether the resulting intensity was bearable. If not the amplitude was incrementally reduced and the final amplitude was used for the rest of the experiment. UCSs were administered before the offset of the stimulus.

Eye Tracking and Fixation Density Maps (FDM)

Eye tracking was done using an Eyelink 1000 Desktop Mount system (SR Research, Ontario Canada) recording the right eye at 1000 Hz. Participants placed their head on a headrest supported under the chin and forehead to keep a stable position. Participants underwent a 13 point calibration / validation procedure at the beginning of each run (1 Baseline run, 1 Conditioning run and 3 runs of Test). The average mean-calibration error across all runs was Mean = 0.36°, Median = .34°, SD = 0.11. 91% of all runs had a calibration better than or equal to .5°.

Fixation events were identified using commonly used parameter definitions [54] (cognitive configuration: saccade velocity threshold = 30° / second, saccade acceleration threshold = 8000° per second^2, motion threshold = .1°). Fixation density maps were computed by spatially smoothing (Gaussian kernel of 1° of FWHM) a 2D histogram of fixation locations, and were transformed to probability densities by normalizing to unit sum. FDMs included the center 500x500 pixels, including all facial elements where fixations were mostly concentrated (~95% of all fixations).

Shock Expectancy Ratings and Autonomic Recordings

After baseline, conditioning and generalization phases, participants rated different faces for subjective shock expectancy by answering the following question, "*How likely is it to receive a shock for this face?*". Faces

were presented in a random order and rated twice. Subjects answered using a 10 steps scale ranging from "*very unlikely*" to "*very likely*" and confirmed by a button press in a self-paced manner.

Electrodermal activity evoked by individual faces was recorded throughout the three phases. Reusable Ag/AgCl electrodes filled with isotonic gel were connected to the palm of the subject's left hand using adhesive collars, placed in thenar/hypothenar configuration. Skin conductance responses were continuously recorded using a Biopac MP100 AD converter and amplifier system at a sampling rate of 500 Hz. Using the Ledalab toolbox [55,56], we decomposed the raw data to phasic and tonic response components after downsampling it to 100 Hz. Ledalab applies a positively constrained deconvolution technique in order to obtain phasic responses for each single trial. We averaged single-trial phasic responses separately for each condition and experimental phase to obtained 18 average values (8 from baseline and test and 2 from the conditioning phase). CS+ trials with UCS were excluded from this analysis. These values were first log-transformed ($\log_{10}(1+SCR)$) and subsequently z-scored for every subject separately (across all conditions and phases) and averaged across subjects. Therefore, negative values indicate phasic response that are smaller than the average responses recorded throughout the experiment. Due to technical problems, SCR data could only be analyzed for n = 51 out of the 61 participants.

Nonlinear Modelling and Model Comparison

We fitted a von Mises function (circular Gaussian) to generalization profiles obtained from subjective ratings, skin-conductance responses and fixation counts at different ROIs by minimizing the likelihood term in (1) following an initial grid search for parameters

$$L(D(x) \mid \theta, \sigma) = \Sigma - \log [N(D(x) - G(x|\theta), 0, \sigma)] \quad (1)$$

where x represents signed angular distances from a given volunteer's CS+ face; $G(x|\theta)$ is a von Mises function defined by the parameter vector θ which codes for the amplitude (difference between peak and base), location (peak position), precision and offset (base value) of the resulting generalization profile; $D(x)$ represents the observed generalization profile for different angular distances; and $N(x| 0, \sigma)$ is the normal probability density function with mean zero and standard deviation of σ. The fitting procedure consisted of finding parameters values that minimized the negative of log transformed probability values. Using log-likelihood ratio test we tested whether this model performed better than a null model consisting of a horizontal line, effectively testing the significance of the additional variance explained by the model.

Classification with Linear Support Vector Machine

We used single trial FDMs for validation of linear support vector machines that were trained to classify exploration patterns obtained during viewing of CS+ and CS– conditions. As the generalization phase had more trials than the baseline phase (3 runs vs. 1 run), we took precautions to make a fair comparison between these phases so that differences between number of trials do not invalidate comparison of accuracies between the baseline and test phases. To this end, we trained and tested a linear SVM classifiers [57] always within a given run, and averaged classification accuracy for the generalization phase across the three runs. We trained a classifier on randomly drawn 50% of the CS + and CS– trials, and tested on the remaining 50% and averaged classification performance across 1000 repetitions using this procedure. To reduce the dimensionality, FDMs were first downscaled 10 times resulting in a vector of 2500 pixels. We further reduced dimensionality by projecting FDMs onto their principal components using two different approaches. We identified the number of *N* principal components corresponding to the elbow where the slope of eigenvalues was levelling off substantially (N = 14, 19, 17, 20 for the 4 runs). We also repeated the analysis simply using N principal components that explained 90% of total variance in each run (N = 63, 69, 74, 75). Both approaches yielded very similar classification results, therefore we report numbers from the first approach in the results section, but present results obtained with both approaches (SFig. 3). Principal components were computed excluding Null trials, UCS trials and oddballs. Loadings on these principal components were scaled using the inverse of the square root of eigenvalues, thus effectively whitening their contributions to the training.

Fixation Similarity Analysis

Fixation similarity analysis was conducted on single participants. Condition specific FDMs (8 faces per baseline and generalization phases) were computed by collecting all fixations across trials on a single map which were then normalized to unit sum. We corrected FDMs by removing the common mean pattern (done separately for baseline and generalization phases). We used 1 - Pearson correlation as the similarity metric. This resulted in a 16x16 similarity matrix per subject. Statistical tests were done after Fisher transformation of correlation values. The multidimensional scaling was conducted on the baseline and generalization phases jointly using the 16x16 similarity matrix as input (*mdscale* in MATLAB). The node coordinates for the baseline condition were linearly transformed to map onto a perfect circle as close as possible using Procrustes mapping, and the same transformation was applied on coordinates of the generalization phase. This has no consequence on MDS results, as proportions between nodes stay unmodified under linear transformations. Importantly, as the similarity metric is extremely sensitive to signal to noise ratio [28] present in the FDMs, we took care that the number of trials between test and baseline phases were exactly the same in order to avoid differences that would have been caused by different signal to noise ratios. To account for unequal number of trials during the baseline (11 repetitions) and test (3 runs x 11 = 33 repetitions) phases, we computed a similarity matrix for each run separately in the generalization phase. These were later averaged across runs for a given participant. This ensured that FDMs of the baseline and generalization phases had similar signal to noise ratios, therefore not favoring the generalization phase for having more trials.

We generated 3 different models based on a quadrature decomposition of a circular similarity matrix. A circular similarity matrix of 8x8 can be obtained using the term $\mathbf{M}\otimes\mathbf{M}$, where $\mathbf{M}$ is a 8x2 matrix in form of

[cos(x) sin(x)], and the operator ⊗ denotes the outer product. x represents angular distances from the CS+ face, is equal to 0 for CS+ and π for CS–. Therefore, while cos(x) is symmetric around the CS+ face, sin(x) is shifted by 90°. For the bottom-up saliency and increased arousal models (Fig. 1B and C) we used M⊗M as a predictor together with a constant intercept. For the tuned exploration model depicted in Figure 1D, we used cos(x) ⊗cos(x) and sin(x)⊗sin(x) to independently model ellipsoid expansion. Together with the intercept this model comprised 3 predictors. Finally the aversive generalization model (Fig. 1E) was created using the predictors of the tuned exploration model in conjunction with a two-dimensional Gaussian centered on the CS+ face (in total 4 predictors). We tested different widths for the Gaussian and took the one that resulted in the best fit. This was equal to 65° of FWHM and similar to the values we observed for explicit ratings and SCR responses.

All linear modeling was conducted using non-redundant and vectorized forms of the symmetric similarity matrices. For a 8x8 similarity matrix this resulted in a vector of 28 entries. To compare different similarity models, we used mixed-effects models where intercept and slope contributed both as fixed- and random-effects (*fitlme* in Matlab). We selected mixed-effect models as these performed much better than models defined uniquely with fixed-effects on intercept and slope. To do model selection we used Bayesian information criteria (BIC) as it compensates for an increase in the number of predictors between different models. For statistical tests on the parameter estimates, these 3 models were fit for every subject individually (*fitlm* in Matlab) and statistical tests were conducted in a pairwise manner using t-test.

Data Sharing

The dataset used in this manuscript has been published as a dataset publication [54]. The code to conduct all the presented analysis, stimuli as well as the figures presented in this manuscript is publicly available [48] and has been developed with Matlab 2016b (MathWorks, Natick MA). This code can be used to download the data set and conduct all analyses presented in this paper.

References

1. Pavlov I. Conditioned reflexes an investigation of the physiological activity of the cerebral cortex. [London]: Oxford University Press Humphrey Milford; 1927.

2. Shepard RN. Toward a universal law of generalization for psychological science. Science. 1987;237: 1317–1323.

3. Dunsmoor JE, Mitroff SR, LaBar KS. Generalization of conditioned fear along a dimension of increasing fear intensity. Learn Mem Cold Spring Harb N. 2009;16: 460–469. doi:10.1101/lm.1431609

4. Kahnt T, Tobler PN. Dopamine regulates stimulus generalization in the human hippocampus. eLife. 2016;5: e12678.

5. Resnik J, Sobel N, Paz R. Auditory aversive learning increases discrimination thresholds. Nat Neurosci. 2011;14: 791–796. doi:10.1038/nn.2802

6. Li W, Howard JD, Parrish TB, Gottfried JA. Aversive Learning Enhances Perceptual and Cortical Discrimination of Indiscriminable Odor Cues. Science. 2008;319: 1842–1845. doi:10.1126/science.1152837

7. Onat S, Büchel C. The neuronal basis of fear generalization in humans. Nat Neurosci. 2015;advance online publication. doi:10.1038/nn.4166

8. Bass MJ, Hull CL. The irradiation of a tactile conditioned reflex in man. J Comp Psychol. 1934;17: 47–65.

9. Tenenbaum JB, Griffiths TL. Generalization, similarity, and Bayesian inference. Behav Brain Sci. 2001;24: 629–640. doi:10.1017/S0140525X01000061

10. Itti L, Koch C. Computational modelling of visual attention. Nat Rev Neurosci. 2001;2: 194–203.

11. Henderson JM. Human gaze control during real-world scene perception. Trends Cogn Sci. 2003;7: 498–504.

12. Yarbus AL. Eye movements and vision. New York: Plenum Press; 1967.

13. Schumann F, Einhäuser-Treyer W, Vockeroth J, Bartl K, Schneider E, König P. Salient features in gaze-aligned recordings of human visual input during free exploration of natural environments. J Vis. 2008;8: 12.1–17. doi:10.1167/8.14.12

14. Peterson MF, Lin J, Zaun I, Kanwisher N. Individual differences in face-looking behavior generalize from the lab to the world. J Vis. 2016;16: 12. doi:10.1167/16.7.12

15. Hayhoe M, Ballard D. Eye movements in natural behavior. Trends Cogn Sci. 2005;9: 188–194. doi:10.1016/j.tics.2005.02.009

16. Dowd EW, Mitroff SR, LaBar KS. Fear generalization gradients in visuospatial attention. Emot Wash DC. 2016;16: 1011–1018. doi:10.1037/emo0000197

17. Cerf M, Frady EP, Koch C. Faces and text attract gaze independent of the task: Experimental data and computer model. J Vis. 2009;9: 10.1-15. doi:10.1167/9.12.10

18. End A, Gamer M. Preferential Processing of Social Features and Their Interplay with Physical Saliency in Complex Naturalistic Scenes. Front Psychol. 2017;8. doi:10.3389/fpsyg.2017.00418

19. Peterson MF, Eckstein MP. Looking just below the eyes is optimal across face recognition tasks. Proc Natl Acad Sci. 2012;109: E3314–E3323. doi:10.1073/pnas.1214269109

20. Jack RE, Garrod OGB, Schyns PG. Dynamic Facial Expressions of Emotion Transmit an Evolving Hierarchy of Signals over Time. Curr Biol. 2014;24: 187–192. doi:10.1016/j.cub.2013.11.064

21. Dunsmoor JE, Prince SE, Murty VP, Kragel PA, LaBar KS. Neurobehavioral mechanisms of human fear generalization. NeuroImage. 2011;55: 1878–1888. doi:10.1016/j.neuroimage.2011.01.041

22. Malcolm GL, Lanyon LJ, Fugard AJB, Barton JJS. Scan patterns during the processing of facial expression versus identity: An exploration of task-driven and stimulus-driven effects. J Vis. 2008;8: 2–2. doi:10.1167/8.8.2

23. Schurgin MW, Nelson J, Iida S, Ohira H, Chiao JY, Franconeri SL. Eye movements during emotion recognition in faces. J Vis. 2014;14: 14–14. doi:10.1167/14.13.14

24. Kriegeskorte N. Representational similarity analysis – connecting the branches of systems neuroscience. Front Syst Neurosci. 2008; 1–28. doi:10.3389/neuro.06.004.2008

25. Mehoudar E, Arizpe J, Baker CI, Yovel G. Faces in the eye of the beholder: Unique and stable eye scanning patterns of individual observers. J Vis. 2014;14: 6.

26. Kanan C, Bseiso DN, Ray NA, Hsiao JH, Cottrell GW. Humans have idiosyncratic and task-specific scanpaths for judging faces. Vision Res. 2015;108: 67–76.

27. Walker-Smith GJ, Gale AG, Findlay JM. Eye movement strategies involved in face perception. Perception. 1977;6: 313–326.

28. Diedrichsen J, Ridgway GR, Friston KJ, Wiestler T. Comparing the similarity and spatial structure of neural representations: A pattern-component model. NeuroImage. 2011;55: 1665–1678. doi:10.1016/j.neuroimage.2011.01.044

29. Masciocchi CM, Mihalas S, Parkhurst D, Niebur E. Everyone knows what is interesting: salient locations which should be fixated. J Vis. 2009;9: 25.1–22. doi:10.1167/9.11.25

30. Dunsmoor JE, Murphy GL. Stimulus Typicality Determines How Broadly Fear Is Generalized. Psychol Sci. 2014;25: 1816–1821. doi:10.1177/0956797614535401

31. Dunsmoor JE, Murphy GL. Categories, concepts, and conditioning: how humans generalize fear. Trends Cogn Sci. 2015;19: 73–77. doi:10.1016/j.tics.2014.12.003

32. Vervliet B, Geens M. Fear generalization in humans: Impact of feature learning on conditioning and extinction. Neurobiol Learn Mem. 2014;113: 143–148. doi:10.1016/j.nlm.2013.10.002

33. Lissek S, Rabin S, Heller RE, Lukenbaugh D, Geraci M, Pine DS, et al. Overgeneralization of conditioned fear as a pathogenic marker of panic disorder. Am J Psychiatry. 2009;167: 47–55. doi:10.1176/appi.ajp.2009.09030410

34. Yue X, Biederman I, Mangini MC, Malsburg C von der, Amir O. Predicting the psychophysical similarity of faces and non-face complex shapes by image-based measures. Vision Res. 2012;55: 41–46. doi:10.1016/j.visres.2011.12.012

35. König P, Wilming N, Kietzmann TC, Ossandón JP, Onat S, Ehinger BV, et al. Eye movements as a window to cognitive processes. J Eye Mov Res. 2016;9. doi:10.16910/jemr.9.5.3

36. Chuk T, Chan AB, Hsiao JH. Understanding eye movements in face recognition using hidden Markov models. J Vis. 2014;14: 8. doi:10.1167/14.11.8

37. Kriegeskorte N, Formisano E, Sorger B, Goebel R. Individual faces elicit distinct response patterns in human anterior temporal cortex. Proc Natl Acad Sci. 2007;104: 20600–20605. doi:10.1073/pnas.0705654104

38. Wiestler T, Diedrichsen J. Skill learning strengthens cortical representations of motor sequences. eLife. 2013;2. doi:10.7554/eLife.00801

39. Cichy RM, Pantazis D, Oliva A. Resolving human object recognition in space and time. Nat Neurosci. 2014;17: 455–462. doi:10.1038/nn.3635

40. Kietzmann TC, Gert AL, Tong F, König P. Representational Dynamics of Facial Viewpoint Encoding. J Cogn Neurosci. 2017;29: 637–651. doi:10.1162/jocn_a_01070

41. Whalen PJ, Rauch SL, Etcoff NL, McInerney SC, Lee MB, Jenike MA. Masked presentations of emotional facial expressions modulate amygdala activity without explicit knowledge. J Neurosci. 1998;18: 411–418.

42. Bannerman RL, Milders M, Gelder B de, Sahraie A. Orienting to threat: faster localization of fearful facial expressions and body postures revealed by saccadic eye movements. Proc R Soc Lond B Biol Sci. 2009;276: 1635–1641. doi:10.1098/rspb.2008.1744

43. Lin JY, Murray SO, Boynton GM. Capture of Attention to Threatening Stimuli without Perceptual Awareness. Curr Biol. 2009;19: 1118–1122. doi:10.1016/j.cub.2009.05.021

44. Lissek S, Biggs AL, Rabin SJ, Cornwell BR, Alvarez RP, Pine DS, et al. Generalization of conditioned fear-potentiated startle in humans: experimental validation and clinical relevance. Behav Res Ther. 2008;46: 678–687. doi:10.1016/j.brat.2008.02.005

45. Cha J, Greenberg T, Carlson JM, DeDora DJ, Hajcak G, Mujica-Parodi LR. Circuit-Wide Structural and Functional Measures Predict Ventromedial Prefrontal Cortex Fear Generalization: Implications for Generalized Anxiety Disorder. J Neurosci. 2014;34: 4043–4053. doi:10.1523/JNEUROSCI.3372-13.2014

46. Greenberg T, Carlson JM, Cha J, Hajcak G, Mujica-Parodi LR. Neural reactivity tracks fear generalization gradients. Biol Psychol. 2013;92: 2–8. doi:10.1016/j.biopsycho.2011.12.007

47. Blakemore C, Campbell FW. On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal images. J Physiol. 1969;203: 237–260.

48. Onat S, Kampermann L. FPSA_FearGen: A Github repository for the analysis and preparation of "Aversive Learning Changes Face-Viewing Strategies—as Revealed by Model-Based Fixation-Pattern Similarity Analysis". [Internet]. 2017. Available: https://github.com/selimonat/FPSA_FearGen.git

49. Brainard DH. The Psychophysics Toolbox. Spat Vis. 1997;10: 433–436.

50. Pelli DG. The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spat Vis. 1997;10: 437–442.

51. Buracas GT, Boynton GM. Efficient design of event-related fMRI experiments using M-sequences. NeuroImage. 2002;16: 801–813.

52. Liu TT, Frank LR. Efficiency, power, and entropy in event-related FMRI with multiple trial types: Part I: theory. NeuroImage. 2004;21: 387–400. doi:10.1016/j.neuroimage.2003.09.030

53. Watson AB, Pelli DG. QUEST: a Bayesian adaptive psychometric method. Percept Psychophys. 1983;33: 113–120.

54. Wilming N, Onat S, Ossandón JP, Açık A, Kietzmann TC, Kaspar K, et al. An extensive dataset of eye movements during viewing of complex images. Sci Data. 2017;4: 160126. doi:10.1038/sdata.2016.126

55. Benedek M, Kaernbach C. Decomposition of skin conductance data by means of nonnegative deconvolution. Psychophysiology. 2010;47: 647–658. doi:10.1111/j.1469-8986.2009.00972.x

56. Benedek M, Kaernbach C. A continuous measure of phasic electrodermal activity. J Neurosci Methods. 2010;190: 80–91. doi:10.1016/j.jneumeth.2010.04.028

57. Chang C-C, Lin C-J. LIBSVM: A library for support vector machines. ACM Trans Intell Syst Technol TIST. 2011;2: 27.

Acknowledgements

The authors wish to thank Nicholas Prins for his input on psychometric estimation, Tim Kitzmann for his input on an early version of this manuscript, Cliodhna Quigley for proof-reading, Patricia Billaudelle and Katrin Harland for their assistance with data collection. This research is supported by the DFG SFB TRR 58.

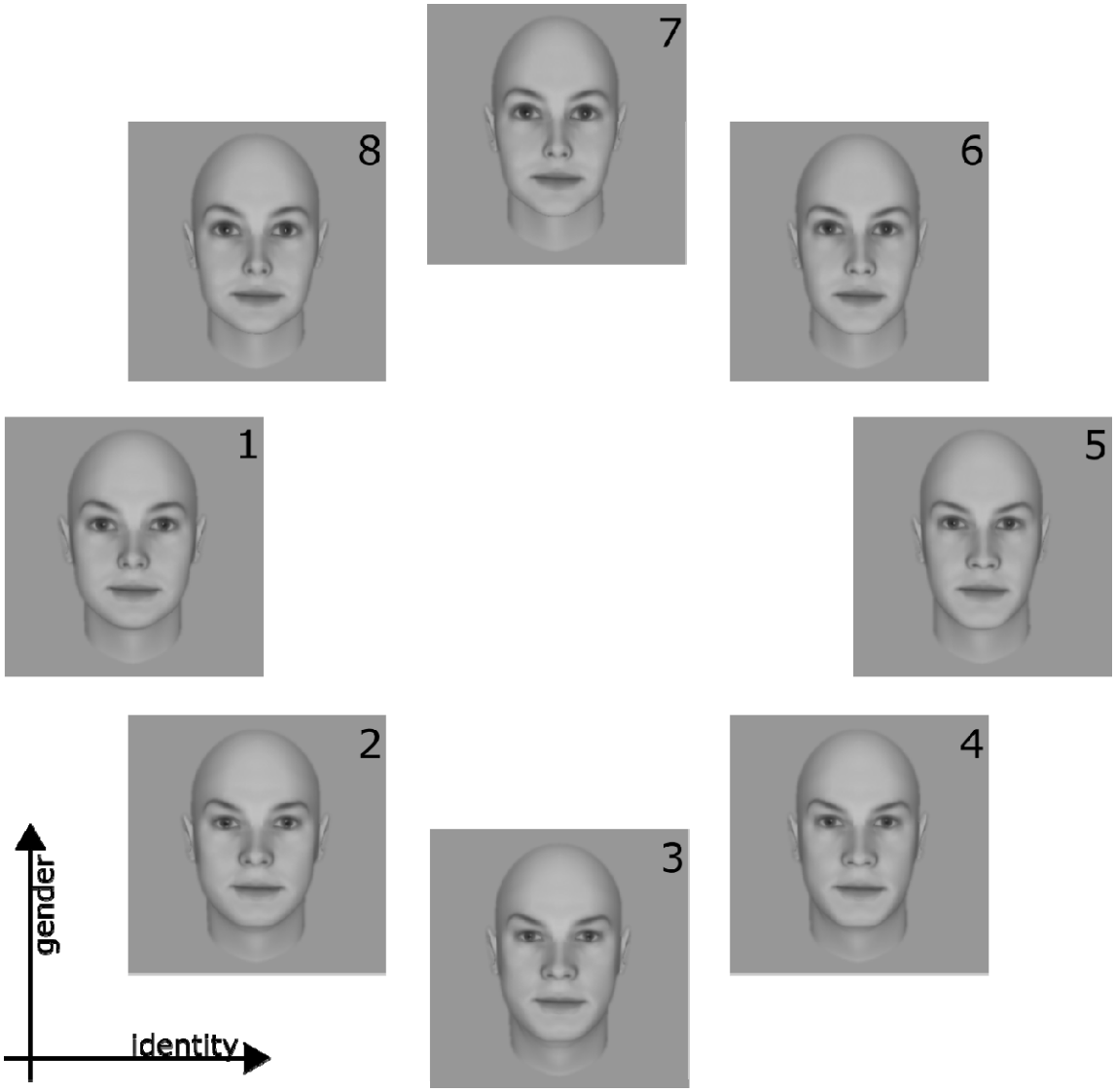
Author contributions

SO supervised the study; SO, CB and LK designed the experiment; LK collected data; SO, LK and NW developed the analysis methods; SO and LK wrote the manuscript; NW, AA, CB critically read the manuscript.

Competing financial interest

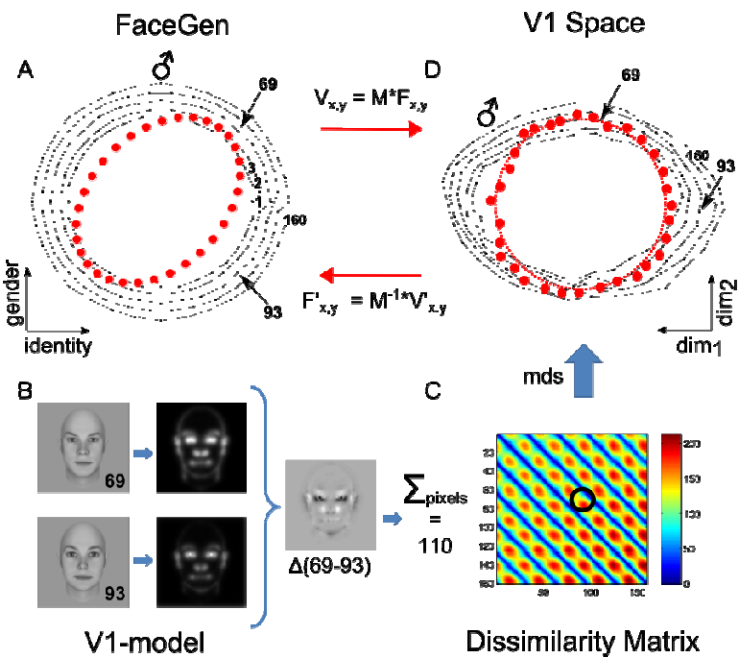
The authors declare no competing financial interests.

Supporting Information



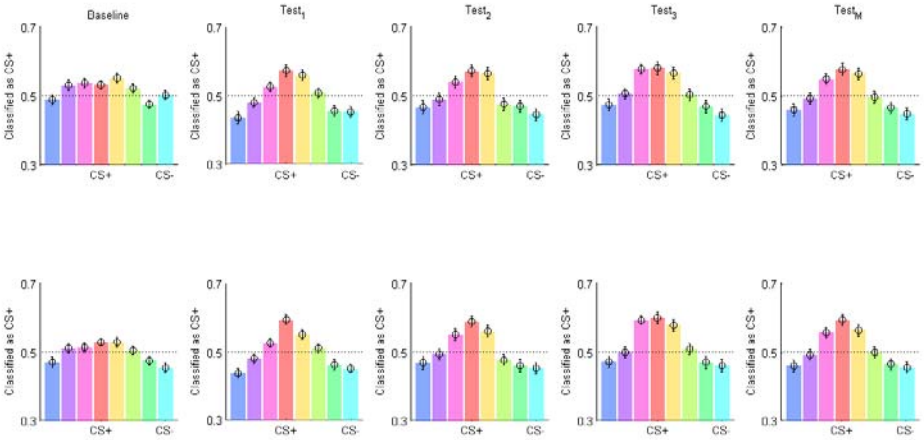
Supplementary Figure 1.

Face Stimuli. Set of 8 faces that were calibrated to form a circular similarity continuum. Faces vary along the two dimensions of gender (vertical axis) and identity (horizontal axis). See SFigure 2 for the calibration process.



Supplementary Figure 2.

Calibration of faces using the V1 model tuned to human psychophysics. (A) Using the FaceGen software, 160 faces forming five concentric circles were generated with coordinates varying in gender and identity dimensions (connected black dots in the left panel). Maximally male faces are located at 12 o'clock direction and indicated with the male symbol. **(B)** V1 representations of faces were modelled according to [34]. This is illustrated for faces 69 and 93. The difference between these two faces resulted in an Euclidean distance of 110. The pair-wise Euclidean distance for all the 160 faces are shown in **(C)** as a dissimilarity matrix. The resulting dissimilarity matrix exhibits 5 major bands corresponding to 5 concentric circles. By applying MDS, we obtained the representational space of V1 shown in (A, right panel). Note that the most male face is 45° counter-clockwise rotated with respect to the main axes of in V1 representation. The mapping between FaceGen coordinates and V1 representational space thus involved a rotation and scaling which was captured by the matrix *M*. We therefore used the inverse of *M*, to achieve coordinates of perfect circularity based on this V1 model. This ensured that faces along the similarity continuum were characterized by controlled changes for every angular step based on the model used.



Supplementary Figure 3 Multivariate Classification of FDMs. Classification accuracies for a linear SVM trained to differentiate between CS+ and CS–. Bars (M ± SEM) show the average proportion of trials of all 8 conditions classified as CS+ (*red*), i.e. correct classification for the actual CS+ condition (*cyan*), and false alarms for CS– trials. Results based on two different dimension reduction methods based on *N* eigenvectors explaining 90% of total variance (upper row) or based on elbow criterium (lower row; see Material and Methods). Baseline and three test runs ($Test_1$, $Test_2$, $Test_3$) are plotted separately. Last column depicts the average across the three test runs. Dotted lines indicate chance level obtained by classification of random label permutation.

Supplementary Table 1. Mixed-effects modeling of the similarity matrices during the baseline phase with the bottom-up model shown in Figure 1B.

Model information:

Number of observations	1708
Fixed effects coefficients	2
Random effects coefficients	122
Covariance parameters	4

Formula:

FPSA_baseline ~ 1 + circle + (1 + circle | subject)

Model fit statistics:

AIC	*BIC*	*LogLikelihood*	*Deviance*
-1682.7	-1650	847.34	-1694.7

Fixed effects coefficients (95% CIs):

Name	*Estimate*	*SE*	*tStat*	*DF*	*pValue*	*Lower*	*Upper*
'(Intercept)'	0.24944	0.0037375	66.741	1706	0	0.24211	0.25677
'circle'	0.063345	0.0079565	7.9614	1706	3.0776e-15	0.047739	0.078951

Random effects covariance parameters (95% CIs):

Group: subject (61 Levels)

Name1	*Name2*	*Type*	*Estimate*	*Lower*	*Upper*
'(Intercept)'	'(Intercept)'	'std'	0.0076928	NaN	NaN
'circle'	'(Intercept)'	'corr'	NaN	NaN	NaN
'circle'	'circle'	'std'	0.044851	NaN	NaN

Group: Error

Name	*Estimate*	*Lower*	*Upper*
'Res Std'	0.14541	NaN	NaN

Supplementary Table 2. Mixed-effects modeling of the similarity matrices during the generalization phase with the arousal model shown in Figure 1C.

Model information:

Number of observations	1708
Fixed effects coefficients	2
Random effects coefficients	122
Covariance parameters	4

Formula:

FPSA_generalization ~ 1 + circle + (1 + circle | subject)

Model fit statistics:

AIC	*BIC*	*LogLikelihood*	*Deviance*
-2748.2	-2715.5	1380.1	-2760.2

Fixed effects coefficients (95% CIs):

Name	*Estimate*	*SE*	*tStat*	*DF*	*pValue*	*Lower*	*Upper*
'(Intercept)'	0.25395	0.0029798	85.222	1706	0	0.24811	0.25979
'circle'	0.091957	0.0099814	9.2129	1706	9.0388e-20	0.07238	0.11153

Random effects covariance parameters (95% CIs):

Group: subject (61 Levels)

Name1	*Name2*	*Type*	*Estimate*	*Lower*	*Upper*
'(Intercept)'	'(Intercept)'	'std'	0.011548	0.0071814	0.01857
'circle'	'(Intercept)'	'corr'	1	NaN	NaN
'circle'	'circle'	'std'	0.071586	0.057996	0.088361

Group: Error

Name	*Estimate*	*Lower*	*Upper*
'Res Std'	0.10434	0.10084	0.10797

Supplementary Table 3. Mixed-effects modeling of the similarity matrices during the baseline phase with the adversity categorization model shown in Fig. 1D.

Model information:

Number of observations	1708
Fixed effects coefficients	3
Random effects coefficients	183
Covariance parameters	7

Formula:

FPSA_baseline ~ 1 + specific + unspecific + (1 + specific + unspecific | subject)

Model fit statistics:

AIC	*BIC*	*LogLikelihood*	*Deviance*
-1764	-1709.6	892.01	-1784

Fixed effects coefficients (95% CIs):

Name	*Estimate*	*SE*	*tStat*	*DF*	*pValue*	*Lower*	*Upper*
'(Intercept)'	0.24944	0.0035811	69.654	1705	0	0.24242	0.25646
'specific'	0.057755	0.013652	4.2304	1705	2.457e-05	0.030977	0.084532
'unspecific'	0.068935	0.010663	6.4647	1705	1.3216e-10	0.048021	0.08985

Random effects covariance parameters (95% CIs):

Group: subject (61 Levels)

Name1	*Name2*	*Type*	*Estimate*	*Lower*	*Upper*
'(Intercept)'	'(Intercept)'	'std'	0.0081486	0.0035081	0.018928
'specific'	'(Intercept)'	'corr'	0.81502	0.80893	0.82093
'unspecific'	'(Intercept)'	'corr'	0.30346	-0.0053757	0.55945
'specific'	'specific'	'std'	0.0913	0.071842	0.11603
'unspecific'	'specific'	'corr'	-0.30479	-0.56025	0.0036179
'unspecific'	'unspecific'	'std'	0.062467	0.045811	0.085177

Group: Error

Name	*Estimate*	*Lower*	*Upper*
'Res Std'	0.13817	0.13344	0.14306

Supplementary Table 4. Mixed-effects modeling of the similarity matrices during the generalization phase with the adversity categorization model shown in Fig. 1D.

Model information:

Number of observations	1708
Fixed effects coefficients	3
Random effects coefficients	183
Covariance parameters	7

Formula:

FPSA_generalization ~ 1 + specific + unspecific + (1 + specific + unspecific | subject)

Model fit statistics:

AIC	*BIC*	*LogLikelihood*	*Deviance*
-2951.7	-2897.3	1485.9	-2971.7

Fixed effects coefficients (95% CIs):

Name	*Estimate*	*SE*	*tStat*	*DF*	*pValue*	*Lower*	*Upper*
'(Intercept)'	0.25395	0.0028004	90.683	1705	0	0.24846	0.25944
'specific'	0.1205	0.014472	8.3259	1705	1.6951e-16	0.09211	0.14888
'unspecific'	0.063418	0.010001	6.3412	1705	2.9123e-10	0.043803	0.083034

Random effects covariance parameters (95% CIs):

Group: subject (61 Levels)

Name1	*Name2*	*Type*	*Estimate*	*Lower*	*Upper*
'(Intercept)'	'(Intercept)'	'std'	0.011778	0.0077913	0.017805
'specific'	'(Intercept)'	'corr'	0.91907	NaN	NaN
'unspecific'	'(Intercept)'	'corr'	0.6935	NaN	NaN
'specific'	'specific'	'std'	0.10647	0.087673	0.12931
'unspecific'	'specific'	'corr'	0.35345	0.34746	0.35941
'unspecific'	'unspecific'	'std'	0.068277	0.054449	0.085617

Group: Error

Name	*Estimate*	*Lower*	*Upper*
'Res Std'	0.09517	0.091916	0.09854

Supplementary Table 5. Mixed-effects modeling of the similarity matrices during the baseline phase with the adversity tuning model shown in Fig. 1E.

Model information:

Number of observations	1708
Fixed effects coefficients	4
Random effects coefficients	244
Covariance parameters	11

Formula:

FPSA_baseline ~ 1 + specific + unspecific + Gaussian + (1 + specific + unspecific + Gaussian | subject)

Model fit statistics:

AIC	*BIC*	*LogLikelihood*	*Deviance*
-1755.5	-1673.9	892.76	-1785.5

Fixed effects coefficients (95% CIs):

Name	*Estimate*	*SE*	*tStat*	*DF*	*pValue*	*Lower*	*Upper*
'(Intercept)'	0.24181	0.046873	5.1589	1704	2.7749e-07	0.14988	0.33374
'specific'	0.057483	0.013761	4.1774	1704	3.0984e-05	0.030494	0.084473
'unspecific'	0.069013	0.010773	6.4063	1704	1.9241e-10	0.047884	0.090142
'Gaussian'	0.015884	0.097484	0.16294	1704	0.87058	-0.17532	0.20709

Random effects covariance parameters (95% CIs):

Group: subject (61 Levels)

Name1	*Name2*	*Type*	*Estimate*	*Lower*	*Upper*
'(Intercept)'	'(Intercept)'	'std'	0.054362	0.022712	0.13012
'specific'	'(Intercept)'	'corr'	0.16475	NaN	NaN
'unspecific'	'(Intercept)'	'corr'	-0.98746	-0.9876	-0.98731
'Gaussian'	'(Intercept)'	'corr'	-0.99227	-0.99237	-0.99216
'specific'	'specific'	'std'	0.091417	0.07191	0.11622
'unspecific'	'specific'	'corr'	-0.31842	-0.32004	-0.31681
'Gaussian'	'specific'	'corr'	-0.041041	NaN	NaN
'unspecific'	'unspecific'	'std'	0.063537	0.046652	0.086532
'Gaussian'	'unspecific'	'corr'	0.96022	NaN	NaN
'Gaussian'	'Gaussian'	'std'	0.12196	0.054259	0.27414

Group: Error

Name	*Estimate*	*Lower*	*Upper*
'Res Std'	0.13805	0.13333	0.14293

Supplementary Table 6. Mixed-effects modeling of the similarity matrices during the generalization phase with the adversity tuning model shown in Fig. 1E.

Model information:

Number of observations	1708
Fixed effects coefficients	4
Random effects coefficients	244
Covariance parameters	11

Formula:

FPSA_generalization ~ 1 + specific + unspecific + Gaussian + (1 + specific + unspecific + Gaussian | subject)

Model fit statistics:

AIC	*BIC*	*LogLikelihood*	*Deviance*
-2945.9	-2864.2	1487.9	-2975.9

Fixed effects coefficients (95% CIs):

Name	*Estimate*	*SE*	*tStat*	*DF*	*pValue*	*Lower*	*Upper*
'(Intercept)'	0.22	0.032419	6.786	1704	1.5856e-11	0.15641	0.28359
'specific'	0.11929	0.014301	8.3414	1704	1.4959e-16	0.091239	0.14734
'unspecific'	0.063765	0.010056	6.3412	1704	2.913e-10	0.044042	0.083488
'Gaussian'	0.070672	0.067933	1.0403	1704	0.29834	-0.062568	0.20391

Random effects covariance parameters (95% CIs):

Group: subject (61 Levels)

Name1	*Name2*	*Type*	*Estimate*	*Lower*	*Upper*
'(Intercept)'	'(Intercept)'	'std'	0.044693	0.010711	0.18649
'specific'	'(Intercept)'	'corr'	-0.89633	-0.89679	-0.89586
'unspecific'	'(Intercept)'	'corr'	-0.73005	-0.73122	-0.72888
'Gaussian'	'(Intercept)'	'corr'	-0.99995	-0.99996	-0.99995
'specific'	'specific'	'std'	0.1047	0.08608	0.12734
'unspecific'	'specific'	'corr'	0.35136	0.35001	0.3527
'Gaussian'	'specific'	'corr'	0.9005	0.90014	0.90086
'unspecific'	'unspecific'	'std'	0.068748	0.054885	0.086113
'Gaussian'	'unspecific'	'corr'	0.72353	0.7229	0.72415
'Gaussian'	'Gaussian'	'std'	0.11757	0.03792	0.36454

Group: Error

Name	*Estimate*	*Lower*	*Upper*
'Res Std'	0.095034	0.091784	0.098399